Tail index estimation, concentration and adaptivity

Stéphane Boucheron, Maud Thomas

arXiv: math.ST/0000000

Tail index estimation, concentration and adaptivity

Stéphane Boucheron, and Maud Thomas

LPMA Université Paris-Diderot,
e-mail: stephane.boucheron@univ-paris-diderot.fr; maud.thomas@univ-paris-diderot.fr

Abstract: This paper presents an adaptive version of the Hill estimator based on Lespki's model selection method. This simple data-driven index selection method is shown to satisfy an oracle inequality and is checked to achieve the lower bound recently derived by Carpentier and Kim. In order to establish the oracle inequality, we derive non-asymptotic variance bounds and concentration inequalities for Hill estimators. These concentration inequalities are derived from Talagrand's concentration inequality for smooth functions of independent exponentially distributed random variables combined with three tools of Extreme Value Theory: the quantile transform, Karamata's representation of slowly varying functions, and Rényi's characterisation of the order statistics of exponential samples. The performance of this computationally and conceptually simple method is illustrated using Monte-Carlo simulations.

60E15, 60G70, 62G30, 60G32.
Keywords and phrases: Hill estimator, adaptivity, Lepskis method, concentration inequalities, order statistics.

1. Introduction

The basic questions faced by Extreme Value Analysis consist in estimating the probability of exceeding a threshold that is larger than the sample maximum and estimating a quantile of an order that is larger than 1 minus the reciprocal of the sample size, that is making inferences on regions that lie outside the support of the empirical distribution. In order to face these challenges in a sensible framework, Extreme Value Theory (EVT) assumes that the sampling distribution F satisfies a regularity condition. Indeed, in heavy-tail analysis, the tail function $\overline{F} = 1-F$ is supposed to be regularly varying that is, $\lim_{\tau\to\infty} \overline{F}(\tau x)/\overline{F}(\tau)$ exists for all $x > 0$. This amounts to assume the existence of some $\gamma > 0$ such that the limit is $x^{-1/\gamma}$ for all x. In other words, if we define the *excess distribution above the threshold* τ by its survival function: $x \mapsto \overline{F}_\tau(x) = \overline{F}(x)/\overline{F}(\tau)$ for $x \geq \tau$, then $\overline{F}$ is regularly varying if and only if F_τ converges weakly towards a Pareto distribution. The sampling distribution F is then said to belong to the *max-domain of attraction* of a Fréchet distribution with index $\gamma > 0$ (abbreviated in $F \in \mathsf{MDA}(\gamma)$) and γ is called the *extreme value index*.

The main impediment to large exceedance and large quantile estimation problems alluded above turns out to be the estimation of the extreme value index.

*Research partially supported by ANR network AMERISKA

Since the inception of Extreme Value Analysis, many estimators have been defined, analysed and implemented into software. Hill [1975] introduced a simple, yet remarkable, collection of estimators: for $k < n$,

$$\widehat{\gamma}(k) = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{X_{(i)}}{X_{(k+1)}} = \frac{1}{k}\sum_{i=1}^{k} i \ln \frac{X_{(i)}}{X_{(i+1)}}$$

where $X_{(1)} \geq \ldots \geq X_{(n)}$ are the *order statistics* of the sample $X_1, \ldots, X_n$ (the non-increasing rearrangement of the sample).

An integer sequence $(k_n)_n$ is said to be *intermediate* if $\lim_{n\to\infty} k_n = \infty$ while $\lim_{n\to\infty} k_n/n = 0$. It is well known that F belongs to $\mathsf{MDA}(\gamma)$ for some $\gamma > 0$ if and only if, for all intermediate sequences $(k_n)_n$, $\widehat{\gamma}(k_n)$ converges in probability towards γ [de Haan and Ferreira, 2006, Mason, 1982]. Under mildly stronger conditions, it can be shown that $\sqrt{k_n}(\widehat{\gamma}(k_n) - \mathbb{E}\widehat{\gamma}(k_n))$ is asymptotically Gaussian with variance γ^2. This suggests that, in order to minimise the quadratic risk $\mathbb{E}[(\widehat{\gamma}(k_n) - \gamma)^2]$ or the absolute risk $\mathbb{E}\,|\widehat{\gamma}(k_n) - \gamma|$, an appropriate choice for k_n has to be made. If k_n is too large, the Hill estimator $\widehat{\gamma}(k_n)$ suffers a large bias and, if k_n is too small, $\widehat{\gamma}(k_n)$ suffers erratic fluctuations. As all estimators of the extreme value index face this dilemma [See Beirlant et al., 2004, de Haan and Ferreira, 2006, Resnick, 2007, and references therein], during the last three decades, a variety of data-driven selection methods for k_n has been proposed in the literature (See Hall and Weissman [1997], Hall and Welsh [1985], Danielsson et al. [2001], Draisma et al. [1999], Drees and Kaufmann [1998], Drees et al. [2000], Grama and Spokoiny [2008], Carpentier and Kim [2014a] to name a few). A related but distinct problem is considered by Carpentier and Kim [2014b]: constructing uniform and adaptive confidence intervals for the extreme value index.

The rationale for investigating adaptive Hill estimation stems from computational simplicity and variance optimality of properly chosen Hill estimators [Beirlant et al., 2006].

In this paper, we combine Talagrand's concentration inequality for smooth functions of independent exponentially distributed random variables (Theorem 2.15) with three traditional tools of EVT: the quantile transform, Karamata's representation for slowly varying functions, and Rényi's characterisation of the joint distribution of order statistics of exponential samples. This allows us to establish concentration inequalities for the Hill process $(\sqrt{k}(\widehat{\gamma}(k) - \mathbb{E}\widehat{\gamma}(k))_k)$ (Theorem 3.3, Propositions 3.9, 3.10 and Corollary 3.13) in Section 3.2. Then in Section 3.3, we build on these concentration inequalities to analyse the performance of a variant of Lepki's rule defined in Sections 2.3 and 3.3: Theorem 3.14 describes an oracle inequality and Corollary 3.18 assesses the performance of this simple selection rule under the assumption that

$$\left|\overline{F}(x)x^{1/\gamma} - C\right| \leq C' x^{\rho/\gamma}$$

for some $\rho < 0$ and $C, C' > 0$. It reveals that the performance of Hill estimators selected by Lepski's method matches known lower bounds. Proofs are given in

Section 4. Finally, in Section 5, we examine the performance of this resulting adaptive Hill estimator for finite sample sizes using Monte-Carlo simulations.

2. Background, notations and tools

2.1. The Hill estimator as a smooth tail statistics

Even though it is possible and natural to characterise the fact that a distribution function F belongs to the max-domain of attraction of a Fréchet distribution with index $\gamma > 0$ by the regular variation property of $\overline{F}$ ($\lim_{\tau\to\infty} \overline{F}(\tau x)/\overline{F}(\tau) = x^{-1/\gamma}$), we will repeatedly use an equivalent characterisation based on the regular variation property of the associated quantile function. We first recall some classical definitions.

If f is a non-decreasing function from (a, b) (where a and b may be infinite) to (c, d), its generalised inverse $f^{\leftarrow} : (c, d) \to (a, b)$ is defined by $f^{\leftarrow}(y) = \inf\{x : a < x < b, f(x) \geq y\}$. The quantile function $F^{\leftarrow}$ is the generalised inverse of the distribution function F. The *tail quantile function* of F is a non-decreasing function defined on $(1, \infty)$ by $U = (1/(1 - F))^{\leftarrow}$, or

$$U(t) = \inf\{x \ : \ F(x) \geq 1 - 1/t\} = F^{\leftarrow}(1 - 1/t) \ .$$

Quantile function plays a prominent role in stochastic analysis thanks to the fact that if Z is uniformly distributed over $[0, 1]$, $F^{\leftarrow}(Z)$ is distributed according to F. In this text, we use a variation of the quantile transform that fits EVT: if E is exponentially distributed, then $U(\exp(E))$ is distributed according to F. Moreover, by the same argument, the order statistics $X_{(1)} \geq \ldots \geq X_{(n)}$ are distributed as a monotone transformation of the order statistics $Y_{(1)} \geq \ldots Y_{(n)}$ of a sample of n independent standard exponential random variables,

$$(X_{(1)}, \ldots, X_{(n)}) \stackrel{d}{=} \left(U(\mathrm{e}^{Y_{(1)}}), \ldots, U(\mathrm{e}^{Y_{(n)}})\right) \ .$$

Thanks to Rényi's representation for order statistics of exponential samples, agreeing on $Y_{(n+1)} = 0$, the rescaled exponential spacings $Y_{(1)} - Y_{(2)}, \ldots, i(Y_{(i)} - Y_{(i+1)}), (n-1)(Y_{(n-1)} - Y_{(n)}), nY_{(n)}$ are independent and exponentially distributed.

The quantile transform and Rényi's representation are complemented by Karamata's representation for slowly varying functions. Recall that a function L is *slowly varying at infinity* if for all $x > 0$, $\lim_{t\to\infty} L(tx)/L(t) = x^0 = 1$. The von Mises condition specifies the form of Karamata's representation [See Resnick, 2007, Corollary 2.1] of the slowly varying component of $U(t)$ ($t^{-\gamma}U(t)$).

Definition 2.1 (von Mises condition). A distribution function F belonging to $\mathsf{MDA}(\gamma), \gamma > 0$, satisfies the von Mises condition if there exist a constant $t_0 \geq 1$, a measurable function η on $(1, \infty)$ and a constant $c = U(t_0)t_0^{-\gamma}$ such that for $t \geq t_0$

$$U(t) = ct^{\gamma} \exp\left(\int_{t_0}^{t} \frac{\eta(s)}{s} \mathrm{d}s\right)$$

with $\lim_{s\to\infty} \eta(s) = 0$. The function η is called the *von Mises function*.

In the sequel, we assume that the sampling distribution $F \in \mathsf{MDA}(\gamma)$, $\gamma > 0$, satisfies the von Mises condition with $t_0 = 1$, von Mises function η and define the non-increasing function $\overline{\eta}$ from $[1, \infty)$ to $[0, \infty)$ by $\overline{\eta}(t) = \sup_{s \geq t} |\eta(s)|$.

Combining the quantile transformation, Rényi's and Karamata's representations, it is straightforward that under the von Mises condition, the sequence of Hill estimators satisfies a distributional identity. It is distributed as a function of the largest order statistics of a standard exponential sample. The next proposition follows easily from the definition of the Hill estimator as a weighted sum of log-spacings, as advocated in [Beirlant et al., 2004].

Proposition 2.2. *The vector of Hill estimators $(\widehat{\gamma}(k))_{k<n}$ is distributed as the random vector*

$$\left(\frac{1}{k} \sum_{i=1}^{k} \int_0^{E_i} \left(\gamma + \eta(\mathrm{e}^{\frac{u}{i} + Y_{(i+1)}}) \right) \mathrm{d}u \right)_{k<n} \tag{2.3}$$

where $(E_1, \ldots, E_n)$ are independent standard exponential random variables while, for $i \leq n$, $Y_{(i)} = \sum_{j=i}^{n} E_j/j$ is distributed like the ith order statistic of an n-sample of the exponential distribution.

For a fixed $k < n$, a second distributional representation is available,

$$\widehat{\gamma}(k) \stackrel{d}{=} \frac{1}{k} \sum_{i=1}^{k} \int_0^{E_i} \left(\gamma + \eta(\mathrm{e}^{u + Y_{(k+1)}}) \right) \mathrm{d}u \tag{2.4}$$

where $(E_1, \ldots, E_k)$ and $Y_{(k+1)}$ are defined as in Proposition 2.2.

This second, simpler, distributional representation stresses the fact that, conditionally on $Y_{(k+1)}$, $\widehat{\gamma}(k)$ is distributed as a mixture of sums of independent identically distributed random variables. Moreover, these independent random variables are close to exponential random variables with scale γ. This distributional identity suggests that the variance of $\widehat{\gamma}(k)$ scales like γ^2/k, an intuition that is corroborated by analysis, see Section 3.1.

The bias of $\widehat{\gamma}(k)$ is connected with the von Mises function η by the next formula

$$\mathbb{E}\widehat{\gamma}(k) - \gamma = \mathbb{E}\left[\int_0^{\infty} \mathrm{e}^{-v} \eta\left(\mathrm{e}^{Y_{(k+1)}} \mathrm{e}^{v} \right) \mathrm{d}v \right] = \mathbb{E}\left[\int_1^{\infty} \frac{\eta\left(\mathrm{e}^{Y_{(k+1)}} v \right)}{v^2} \mathrm{d}v \right] .$$

Henceforth, let b be defined for $t > 1$ by

$$b(t) = \int_1^{\infty} \frac{\eta(tv)}{v^2} \mathrm{d}v = t \int_t^{\infty} \frac{\eta(v)}{v^2} \mathrm{d}v \,. \tag{2.5}$$

The quantity $b(t)$ is the bias of the Hill estimator $\widehat{\gamma}(k)$ given $\overline{F}(X_{(k+1)}) = 1/t$. The second expression for b shows that b is differentiable with respect to t (even though η might be nowhere differentiable), and that

$$b'(t) = \frac{b(t) - \eta(t)}{t} \,.$$

The von Mises function governs both the rate of convergence of $U(tx)/U(t)$ towards x^γ, or equivalently of $\overline{F}(tx)/\overline{F}(t)$ towards $x^{-1/\gamma}$, and the rate of convergence of $|\mathbb{E}\widehat{\gamma}(k) - \gamma|$ towards 0.

In the sequel, we work on the probability space where the independent standard exponential random variables $E_i, 1 \leq i \leq n$ are all defined, and therefore consider the Hill estimators defined by Representation (2.3).

2.2. Frameworks

The difficulty in extreme value index estimation stems from the fact that, for any collection of estimators, for any intermediate sequence $(k_n)_n$, and for any $\gamma > 0$, there is a distribution function $F \in \mathsf{MDA}(\gamma)$ such that the bias $|\mathbb{E}\widehat{\gamma}(k_n) - \gamma|$ decays at an arbitrarily slow rate. This has led authors to put conditions on the rate of convergence of $U(tx)/U(t)$ towards x^γ as t tends to infinity while $x > 0$, or equivalently on the rate of convergence of $\overline{F}(tx)/\overline{F}(t)$ towards $x^{-1/\gamma}$. These conditions have then to be translated into conditions on the rate of decay of the bias of estimators. As we focus on Hill estimators, the connection between the rate of convergence of $U(tx)/U(t)$ towards x^γ and the rate of decay of the bias is transparent and well-understood [Segers, 2002]: the theory of O-regular variation provides an adequate setting for describing both rates of convergence [Bingham et al., 1987]. In words, if a positive function g defined over $[1, \infty)$ is such that for some $\alpha \in \mathbb{R}$, for all $\Lambda > 1$, $\limsup_t \sup_{x\in[1,\Lambda]} g(tx)/g(t) < \infty$, g is said to have *bounded increase*. If g has bounded increase, the class $O\Pi_g$ is the class of measurable functions f on some interval $[a, \infty), a > 0$, such that as $t \to \infty$, $f(tx) - f(t) = O(g(t))$ for all $x \geq 1$.

For example, the analysis carried out by Carpentier and Kim [2014a] rests on the condition that, if $F \in \mathsf{MDA}(\gamma)$, for some $C > 0$, $D \neq 0$ and $\rho < 0$,

$$\left| \frac{\overline{F}(x)}{x^{-1/\gamma}} - C \right| \leq D x^{\rho/\gamma} \,. \tag{2.6}$$

This condition implies that $\ln(t^{-\gamma}U(t)) \in O\Pi_g$ with $g(t) = t^\rho$ [Segers, 2002, p. 473]. Thus under the von Mises condition, Condition (2.6) implies that the function $\int_t^\infty (\eta(s)/s)\mathrm{d}s$ belongs to $O\Pi_g$ with $g(t) = t^\rho$. Moreover, the Abelian and Tauberian Theorems from [Segers, 2002] assert that $\left|\int_t^\infty (\eta(s)/s)\mathrm{d}s\right| \in O\Pi_g$ if and only if $|\mathbb{E}\widehat{\gamma}(k_n) - \gamma| = O(g(n/k_n))$ for any intermediate sequence $(k_n)_n$.

In this text, we are ready to assume that if $F \in \mathsf{MDA}(\gamma)$, then for some $C > 0$ and $\rho < 0$,

$$|\mathbb{E}\widehat{\gamma}(k_n) - \gamma| \leq C \left(\frac{n}{k_n}\right)^\rho \,.$$

However, we do not want to assume that U (or equivalently $\overline{F}$) satisfies a so-called second-order regular variation property ($t \mapsto |x^{-\gamma}U(tx)/U(t) - 1|$ is asymptotically equivalent to a ρ-regularly varying function where $\rho < 0$). By [Segers, 2002], this would be equivalent to assuming that $t \mapsto |b(t)|$ is ρ-regularly varying.

Indeed, assuming as in [Hall and Welsh, 1985] and several subsequent papers that F satisfies

$$\overline{F}(x) = Cx^{-1/\gamma}\left(1 + Dx^{\rho/\gamma} + o(x^{\rho/\gamma})\right) \tag{2.7}$$

where $C > 0, D \neq 0$ are constants and $\rho < 0$, or equivalently [Csörgő, Deheuvels, and Mason, 1985, Drees and Kaufmann, 1998] that U satisfies

$$U(t) = C^{\gamma} t^{\gamma}\left(1 + \gamma D C^{\rho} t^{\rho} + o(t^{\rho})\right)$$

makes the problem of extreme value index estimation easier (but not easy). These conditions entail that, for any intermediate sequence (k_n), the ratio $|\mathbb{E}[\widehat{\gamma}(k_n) - \gamma]|/(n/k_n)^{\rho}$ converges towards a finite limit as n tends to ∞ [Beirlant et al., 2004, de Haan and Ferreira, 2006, Segers, 2002], this makes the estimation of the second-order parameter a very natural intermediate objective [See for example Drees and Kaufmann, 1998].

2.3. Lepski's method and adaptive tail index estimation

The necessity of developing data-driven index selection methods is illustrated in Figure 1, which displays the estimated standardised root mean squared error (RMSE) of Hill estimators

$$\mathbb{E}\left[\left(\frac{\widehat{\gamma}(k)}{\gamma} - 1\right)^2\right]^{1/2}$$

as a function of k for four related sampling distributions which all satisfy the second-order condition (2.7).

Under this second-order condition (2.7), Hall and Welsh proved that the asymptotic mean squared error of the Hill estimator is minimal for sequences $(k_n^*)_n$ satisfying

$$k_n^* \sim \left(\frac{C^{2|\rho|}(1+|\rho|)^2}{2D^2|\rho|^3}\right)^{1/(1+2|\rho|)} n^{2|\rho|/(1+2|\rho|)} \ .$$

Since, $C > 0$, $D \neq 0$ and the second-order parameter $\rho < 0$ are usually unknown, many authors have been interested in the construction of data-driven selection procedure for k_n under conditions such as (2.7), a great deal of ingenuity has been dedicated to the estimation of the second-order parameters and to the use of such estimates when estimating first order parameters.

As we do not want to assume a second-order condition such as (2.7), we resort to Lepski's method which is a general attempt to balance bias and variance.

Since its introduction [Lepski, 1991], this general method for model selection that has been proved to achieve adaptivity and provide one with oracle inequalities in a variety of inferential contexts ranging from density estimation to inverse problems and classification [Lepski, 1990, 1991, 1992, Lepski and

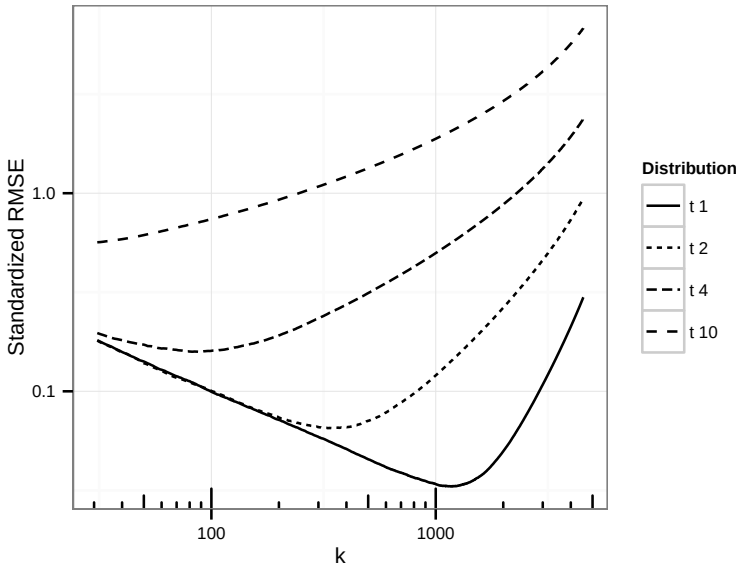

FIG 1. *Estimated standardised* RMSE *as a function of* k *for samples of size* 10000 *from Student's distributions with different degrees of freedom* $\nu = 1, 2, 4, 10$. *All four distributions satisfy condition* (2.7) *with* $|\rho| = 2/\nu$. *The increasing parts of the lines reflect the values of* ρ. RMSE *is estimated by averaging over* 5000 *Monte-Carlo simulations.*

Tsybakov, 2000]. Very readable introductions to Lepski's method and its connections with penalised contrast methods can be found in [Birgé, 2001, Mathé, 2006]. In Extreme Value Theory, we are aware of three papers that explicitly rely on this methodology: [Drees and Kaufmann, 1998], [Grama and Spokoiny, 2008] and [Carpentier and Kim, 2014a].

The selection rule analysed in the present paper (see Section 3.3 for a precise definition) is a variant of the preliminary selection rule introduced in [Drees and Kaufmann, 1998]

$$\overline{\kappa}_n(r_n) = \min\left\{k \in \{2,\dots,n\} \colon \max_{2\le i\le k} \sqrt{i}|\widehat{\gamma}(i) - \widehat{\gamma}(k)| > r_n\right\} \tag{2.8}$$

where $(r_n)_n$ is a sequence of thresholds such that $\sqrt{\ln\ln n} = o(r_n)$ and $r_n = o(\sqrt{n})$, and $\widehat{\gamma}(i)$ is the Hill estimator computed from the $(i+1)$ largest order statistics. The definition of this "stopping time" is motivated by Lemma 1 from [Drees and Kaufmann, 1998] which asserts that, under the von Mises condition,

$$\max_{2\le i\le k_n} \sqrt{i}|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i)\right]| = O_P\left(\sqrt{\log\log n}\right) \quad .$$

In words, this selection rule almost picks out the largest index k such that, for all i smaller than k, $\widehat{\gamma}(k)$ differs from $\widehat{\gamma}(i)$ by a quantity that is not much larger

than the typical fluctuations of $\widehat{\gamma}(i)$. This index selection rule can be performed graphically by interpreting an alternative Hill plot as shown on Figure 2 [See Drees et al., 2000, Resnick, 2007, for a discussion on the merits of alt-Hill plots].

Under mild conditions on the sampling distribution, $\overline{\kappa}_n(r_n)$ should be asymptotically equivalent to the deterministic sequence

$$\widetilde{\kappa}_n(r_n) = \min\left\{k \in \{2,\dots,n\} \colon \max_{2\le i\le k} \sqrt{i}\,|\mathbb{E}\,[\widehat{\gamma}(i) - \widehat{\gamma}(k)]| > r_n\right\}.$$

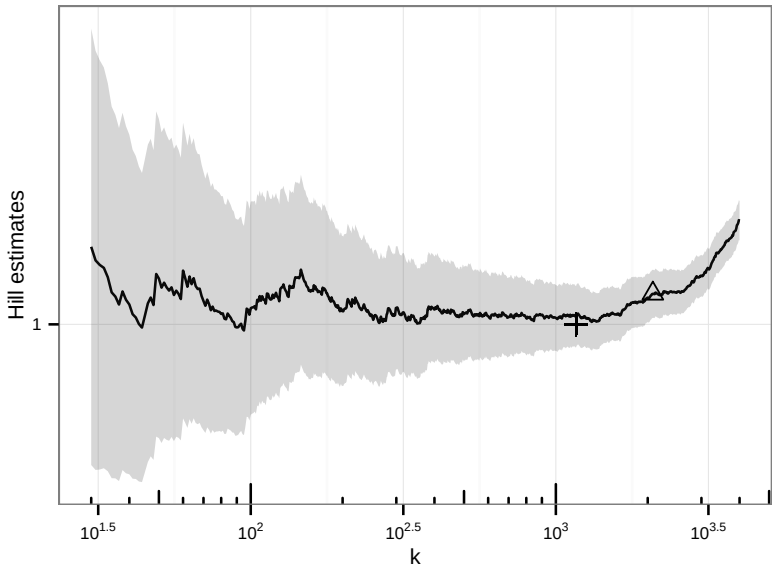

FIG 2. *Lepski's method illustrated on a alt-Hill plot. The plain line describes the sequence of Hill estimates as a function of index k computed on a pseudo-random sample of size $n = 10000$ from Student distribution with 1 degree of freedom (Cauchy distribution). Hill estimators are computed from the positive order statistics. The grey ribbon around the plain line provides a graphic illustration of Lepski's method. For a given value of i, the width of the ribbon is $2r_n\widehat{\gamma}(i)/\sqrt{i}$. A point $(k,\widehat{\gamma}(k))$ on the plain line corresponds to an eligible index if the horizontal segment between this point and the vertical axis lies inside the ribbon that is, if for all $i, 30 \le i < k$, $|\widehat{\gamma}(k) - \widehat{\gamma}(i)| \le r_n\widehat{\gamma}(i)/\sqrt{i}$. If r_n were replaced by an appropriate quantile of the Gaussian distribution, the grey ribbon would just represent the confidence tube that is usually added on Hill plots. The triangle represents the selected index with $r_n = \sqrt{2.1\ln\ln n}$. The cross represents the oracle index estimated from Monte-Carlo simulations, see Table 2.*

The intuition behind the definition of $\overline{\kappa}_n(r_n)$ is the following: if the bias is increasing with index i, and if the bias suffered by $\widehat{\gamma}(k)$ is smaller than the typical fluctuations of $\widehat{\gamma}(k)$, then the index k should be eligible that is, should pass all the pairwise tests with high probability.

The goal of Drees and Kaufmann [1998] was not to investigate the performance of the preliminary selection rule defined in Display (2.8) but to design

a selection rule $\widehat{\kappa}_n(r_n)$, based on $\overline{\kappa}_n(r_n)$, that would, under second-order conditions, asymptotically mimic the optimal selection rule k_n^*. Some of their intermediate results shed light on the behaviour of $\overline{k}_n(r_n)$ for a wide variety of choices for r_n. As they are relevant to our work, we will briefly review them.

Drees and Kaufmann [1998] characterise the asymptotic behaviours of $\widetilde{\kappa}_n(r_n)$ and $\widehat{\kappa}_n(r_n)$ when (r_n) grows sufficiently fast that is, $\sqrt{\ln \ln n} = o(r_n)$ and $r_n = o(\sqrt{n})$. Indeed, Corollary 1 asserts that if $|b(t)| \sim Ct^\rho$, then the sequence $\widetilde{\kappa}_n(r_n)$ satisfies

$$\frac{\widetilde{\kappa}_n(r_n)}{k_n^*} \sim c_\rho r_n^{2/(1+2|\rho|)}$$

where c_ρ is a constant depending on ρ, and that

$$\frac{\widehat{\kappa}_n(r_n)}{\widetilde{\kappa}_n(r_n)} \underset{n\to\infty}{\longrightarrow} 1 \quad \text{in probability}$$

so that

$$r_n^{-2/(1+2|\rho|)} \frac{\widehat{\kappa}_n(r_n)}{k_n^*} \underset{n\to\infty}{\longrightarrow} c_\rho \quad \text{in probability.}$$

This suggests that using $\widehat{\kappa}_n(r_n)$ instead of k_n^* has a price of the order $r_n^{2/(1+2|\rho|)}$.

Not too surprisingly, Corollary 1 from [Drees and Kaufmann, 1998] implies that the preliminary selection rule tends to favor smaller variance over reduced bias.

Our goal, as in [Carpentier and Kim, 2014a, Grama and Spokoiny, 2008], is to derive non-asymptotic risk bounds. We briefly review their approaches. Both papers consider sequences of estimators $\widehat{\gamma}(1), \ldots, \widehat{\gamma}(k), \ldots$ defined by thresholds $\tau_1 \leq \ldots \leq \tau_k \leq \ldots$. For each i, the estimator $\widehat{\gamma}(i)$ is computed from sample points that exceed τ_i if there are any. For example, in [Carpentier and Kim, 2014a], $\tau_k = \theta^k$ for some $\theta > 1$ and $1/\widehat{\gamma}(k) = \ln \overline{F}_n(\tau_k) - \ln \overline{F}_n(\tau_{k+1})$. Given a sample, an estimator $\widehat{\gamma}(k)$ is considered eligible, if for all $i \geq k$ such that $\overline{F}(\tau_i)$ is not too small, $|\widehat{\gamma}(i) - \widehat{\gamma}(k)|$ is smaller than a random quantity that is supposed to bound the typical fluctuations of $\widehat{\gamma}(i)$. The selected index $\widehat{k}$ is the largest eligible index. In both papers, the rationale for working with some special collection of estimators seems to be the ability to derive non-asymptotic deviation inequalities for $\widehat{\gamma}(k)$ either from exponential inequalities for log-likelihood ratio statistics or from simple binomial tail inequalities such as Bernstein's inequality [See Boucheron et al., 2013, Section 2.8].

In models satisfying Condition (2.7), the estimators from [Grama and Spokoiny, 2008] achieve the optimal rate up to a $\log(n)$ factor. Carpentier and Kim [2014a] prove that the risk of their data-driven estimator decays at the optimal rate $n^{|\rho|/(1+2|\rho|)}$ up to a factor $r_n^{2|\rho|/(1+2|\rho|)} = (\ln \ln n)^{|\rho|/(1+2|\rho|)}$ in models satisfying Condition (2.6).

We aim at achieving optimal risk bounds under Condition (2.6), using a simple estimation method requiring almost no calibration effort and based on mainstream extreme value index estimators. Before describing the keystone of our approach in Section 2.5, we recall the recent lower risk bound for adaptive extreme value index estimation.

2.4. Lower bound

One of key results in [Carpentier and Kim, 2014a] is a lower bound on the accuracy of adaptive tail index estimation. This lower bound reveals that, just as for estimating a density at a point [Lepski, 1991, 1992], as far as tail index estimation is concerned, adaptivity has a price. Using Fano's Lemma, and a Bayesian game that extends cleanly in the frameworks of [Grama and Spokoiny, 2008] and [Novak, 2014], Carpentier and Kim were able to prove the next minimax lower bound.

Theorem 2.9. *Let $\rho_0 < -1$, and $0 \leq v \leq \mathrm{e}/(1+2\mathrm{e})$. Then, for any tail index estimator $\widehat{\gamma}$ and any sample size n such that $M = \lfloor \ln n \rfloor > \mathrm{e}/v$, there exists a probability distribution P such that*

i) $P \in \mathsf{MDA}(\gamma)$ *with* $\gamma > 0$,

ii) P *meets the von Mises condition with von Mises function* η *satisfying*

$$\overline{\eta}(t) \leq \gamma t^{\rho}$$

for some $\rho_0 \leq \rho < 0$,

iii)

$$P\left\{ |\widehat{\gamma} - \gamma| \geq \frac{C_\rho}{4}\gamma \left(\frac{v \ln\ln n}{n} \right)^{|\rho|/(1+2|\rho|)} \right\} \geq \frac{1}{1+2\mathrm{e}}$$

and

$$\mathbb{E}_P\left[\frac{|\widehat{\gamma} - \gamma|}{\gamma} \right] \geq \frac{C_\rho}{4(1+2\mathrm{e})} \left(\frac{v \ln\ln n}{n} \right)^{|\rho|/(1+2|\rho|)},$$

with $C_\rho = 1 - \exp\left(-\frac{1}{2(1+2|\rho|)^2} \right)$.

Using Birgé's Lemma instead of Fano's Lemma, we provide a simpler, shorter proof of this theorem (Appendix D).

The lower rate of convergence provided by Theorem 2.9 is another incentive to revisit the preliminary tail index estimator from [Drees and Kaufmann, 1998], but, instead of using a sequence $(r_n)_n$ of order larger than $\sqrt{\ln\ln n}$ in order to calibrate pairwise tests and ultimately to design estimators of the second-order parameter (if there are any), it is worth investigating a minimal sequence where r_n is of order $\sqrt{\ln\ln n}$, and check whether the corresponding adaptive estimator achieves the Carpentier-Kim lower bound (Theorem 2.9).

In this paper, we focus on r_n of the order $\sqrt{\ln\ln n}$. The rationale for imposing r_n of the order $\sqrt{\ln\ln n}$ can be understood by the fact that if $\limsup r_n/(\gamma\sqrt{2\ln\ln n}) < 1$, even if the sampling distribution is a pure Pareto distribution with shape parameter γ ($\overline{F}(x) = (x/\tau)^{-1/\gamma}$ for $x \geq \tau > 0$), the preliminary selection rule will, with high probability, select a small value of k and thus pick out a suboptimal estimator. This can be justified using results from [Darling and Erdös, 1956] (See Appendix A for details).

Such an endeavour requires sharp probabilistic tools. They are the topic of the next section.

2.5. Talagrand's concentration phenomenon for products of exponential distributions

Before introducing Talagrand's Theorem, which will be the key tool of our investigation, we comment and motivate the use of concentration arguments in Extreme Value Theory. Talagrand's concentration phenomenon for products of exponential distributions is one instance of a general phenomenon: concentration of measure in product spaces [Ledoux, 2001, Ledoux and Talagrand, 1991]. The phenomenon may be summarised in a simple quote: functions of independent random variables that do not depend too much on any of them are almost constant [Talagrand, 1996a].

This quote raises a first question: in which way are tail functionals (as used in Extreme Value Theory) smooth functions of independent random variables? We do not attempt here to revisit the asymptotic approach described by [Drees, 1998b] which equates smoothness with Hadamard differentiability. Our approach is non-asymptotic and our conception of smoothness somewhat circular, smooth functionals are these functionals for which we can obtain good concentration inequalities.

The concentration approach helps to split the investigation in two steps: the first step consists in bounding the fluctuations of the random variable under concern around its median or its expectation, while the second step focuses on the expectation. This approach has serioulsy simplified the investigation of suprema of empirical processes and made the life of many statisticians easier [Koltchinskii, 2008, Massart, 2007, Talagrand, 1996b, 2005]. Up to our knowledge, the impact of the concentration of measure phenomenon in Extreme Value Theory has received little attention. To point out the potential uses of concentration inequalities in the field of Extreme Value Theory is one purpose of this paper. In statistics, concentration inequalities have proved very useful when dealing with adaptivity issues: sharp, non-asymptotic tail bounds can be combined with simple union bounds in order to obtain uniform guarantees of the risk of collection of estimators. Using concentration inequalities to investigate adaptive choice of the number of order statistics to be used in tail index estimation is a natural thing to do.

Deriving authentic concentration inequalities for Hill estimators is not straightforward. Fortunately, the construction of such inequalities turns out to be possible thanks to general functional inequalities that hold for functions of independent exponentially distributed random variables. We recall these inequalities (Proposition 2.10 and Theorem 2.15) that have been largely overlooked in statistics. A thorough and readable presentation of these inequalities can be found in [Ledoux, 2001]. We start by the easiest result, a variance bound that pertains to the family of Poincaré inequalities.

Proposition 2.10 (Poincaré inequality for exponentials, [Bobkov and Ledoux, 1997])**.** *If f is a differentiable function over $\mathbb{R}^n$, and $Z = f(E_1, \ldots, E_n)$ where $E_1, \ldots, E_n$ are independent standard exponential random variables, then*

$$\mathrm{Var}(Z) \leq 4\mathbb{E}\left[\|\nabla f\|^2\right] .$$

Remark 2.11. The constant 4 can not be improved.

The next corollary is stated in order to point the relevance of this Poincaré inequality to the analysis of general order statistics and their functionals. Recall that the *hazard rate* of an absolutely continuous probability distribution with distribution F is: $h = f/\overline{F}$ where f and $\overline{F} = 1 - F$ are the density and the survival function associated with F, respectively.

Corollary 2.12. *Assume the distribution of X has a positive density, then the kth order statistic $X_{(k)}$ satisfies*

$$\mathrm{Var}(X_{(k)}) \leq C \sum_{i=k}^{n} \frac{1}{i^2} \mathbb{E}\left[\frac{1}{h(X_{(k)})^2}\right] \leq \frac{C}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\frac{1}{h(X_{(k)})^2}\right]$$

where C can be chosen as 4.

Remark 2.13. By Smirnov's Lemma [de Haan and Ferreira, 2006], C can not be smaller than 1. If the distribution of X has a non-decreasing hazard rate, the factor of 4 can be improved into a factor 2 [Boucheron and Thomas, 2012].

Bobkov and Ledoux [1997], Maurey [1991], Talagrand [1991] show that smooth functions of independent exponential random variables satisfy Bernstein type concentration inequalities. The next result is extracted from the derivation of Talagrand's concentration phenomenon for product of exponential random variables in [Bobkov and Ledoux, 1997].

The definition of sub-gamma random variables will be used in the formulation of the theorem and in many arguments.

Definition 2.14. A real-valued centered random variable X is said to be *sub-gamma* on the right tail with variance factor v and scale parameter c if

$$\ln \mathbb{E}\mathrm{e}^{\lambda X} \leq \frac{\lambda^2 v}{2(1-c\lambda)} \text{ for every } \lambda \quad \text{such that} \quad 0 < \lambda < 1/c\,.$$

We denote the collection of such random variables by $\Gamma_+(v,c)$. Similarly, X is said to be sub-gamma on the left tail with variance factor v and scale parameter c if $-X$ is sub-gamma on the right tail with variance factor v and tail parameter c. We denote the collection of such random variables by $\Gamma_-(v,c)$.

If $X - \mathbb{E}X \in \Gamma_+(v,c)$, then for all $\delta \in (0,1)$, then with probability larger than $1-\delta$,

$$X \leq \mathbb{E}X + \sqrt{2v\ln(1/\delta)} + c\ln(1/\delta)\ .$$

Theorem 2.15. *If f is a differentiable function on $\mathbb{R}^n$ with $\max_i |\partial_i f| < \infty$, and $Z = f(E_1,\ldots,E_n)$ where $E_1,\ldots,E_n$ are n independent standard exponential random variables. Let $c < 1$, then for all λ such that $0 \leq \lambda \max_i |\partial_i f| \leq c$,*

$$\mathrm{Ent}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{2\lambda^2}{1-c}\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\|\nabla f\|^2\right]\ .$$

Let v be the essential supremum of $\|\nabla f\|^2$, then Z is sub-gamma on both tails with variance factor $4v$ and scale factor $\max_i |\partial_i f|$.

Again, we illustrate the relevance of these versatile tools to the analysis of general order statistics. This general theorem implies that if the sampling distribution has non-decreasing hazard rate, then the order statistics $X_{(k)}$ satisfy Bernstein type inequalities [see Boucheron et al., 2013, Section 2.8] with variance factor $4/k\mathbb{E}\left[1/h(X_{(k)})^2\right]$ (the Poincaré estimate of variance), and scale parameter $(\sup_x 1/h(x))/k$. Starting back from the Efron-Stein-Steele inequality, the authors derived a somewhat sharper inequality [Boucheron and Thomas, 2012].

Corollary 2.16. *Assume the distribution function F has non-decreasing hazard rate h that is, $U \circ \exp$ is C^1 and concave. Let $Z = f(E_1, \ldots, E_n) = (U \circ \exp)\left(\sum_{i=k}^n E_i/i\right)$ be distributed as the kth order statistic of a sample distributed according to F.*

Then Z is sub-gamma on both tails with variance factor $4/k\,(1+1/k)\,\mathbb{E}[1/h(Z)^2]$ and scale factor $1/(k \inf_x h(x))$.

This corollary describes in which way central, intermediate and extreme order statistics can be portrayed as smooth functions of independent exponential random variables. This possibility should not be taken for granted as it is non trivial to capture in a non-asymptotic way the tail behaviour of maxima of independent Gaussians [Boucheron and Thomas, 2012, Chatterjee, 2014, Ledoux, 2001]. In the next section, we show in which way the Hill estimator can fit into this picture.

3. Main results

In this section, the sampling distribution F is assumed to belong to $\mathsf{MDA}(\gamma)$ with $\gamma > 0$ and to satisfy the von Mises condition (Definition 2.1) with von Mises function η.

3.1. Bounding the variance of the Hill estimator

It is well known, that under the von Mises condition, if $(k_n)_n$ is an intermediate sequence, the sequence $\sqrt{k_n}\left(\widehat{\gamma}(k_n) - \mathbb{E}\widehat{\gamma}(k_n)\right)$ converges in distribution towards $\mathcal{N}(0,\gamma^2)$ suggesting that the variance of $\widehat{\gamma}(k_n)$ scales like γ^2/k_n [See Beirlant et al., 2004, de Haan and Ferreira, 2006, Geluk et al., 1997, Resnick, 2007].

In this subsection, we will use the representation (2.4):

$$\widehat{\gamma}(k) \stackrel{d}{=} \frac{1}{k}\sum_{i=1}^{k} \int_0^{E_i} \left(\gamma + \eta(\mathrm{e}^{u+Y_{(k+1)}})\right) \mathrm{d}u$$

Proposition 3.1 provides us with a handy upper bound on $\mathrm{Var}[\widehat{\gamma}(k)] - \gamma^2/k$ using the von Mises function.

Proposition 3.1. *Let $\widehat{\gamma}(k)$ be the Hill estimator computed from the $(k+1)$ largest order statistics of an n-sample from F. Then,*

$$-\frac{2\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)\right] \le \mathrm{Var}[\widehat{\gamma}(k)] - \frac{\gamma^2}{k} \le \frac{2\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)\right] + \frac{5}{k}\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)^2\right].$$

The next Abelian result might help in appreciating these variance bounds.

Proposition 3.2. *Assuming that η is ρ-regular varying with $\rho < 0$, then for any intermediate sequence $(k_n)_n$,*

$$\lim_{n\to\infty} \frac{k_n \operatorname{Var}(\widehat{\gamma}(k_n)) - \gamma^2}{\eta(n/k_n)} = \frac{2\gamma}{(1-\rho)^2}\,.$$

We may now move to genuine concentration inequalities for the Hill estimator.

3.2. Concentration inequalities for the Hill estimators

The exponential representation (2.3) suggests that the Hill estimator $\widehat{\gamma}(k)$ should be approximately distributed according to a gamma(k,γ) distribution where k is the shape parameter and γ the scale parameter. We expect the Hill estimators to satisfy Bernstein type concentration inequalities that is, to be sub-gamma on both tails with variance factors connected to the tail index γ and to the von Mises function. Representation (2.3) actually suggests more. Following [Drees and Kaufmann, 1998], we actually expect the sequence $\sqrt{k}(\widehat{\gamma}(k) - \mathbb{E}\widehat{\gamma}(k))$ to behave like normalized partial sums of independent square integrable random variables that is, we believe $\max_{2\le k\le n} \sqrt{k}(\widehat{\gamma}(k) - \mathbb{E}\widehat{\gamma}(k))$ to scale like $\sqrt{\ln\ln n}$ and to be sub-gamma on both tails. The purpose of this section is to meet these expectations in a non-asymptotic way.

Proofs use the Markov property of order statistics: conditionally on the $(J+1)$th order statistic, the first largest J order statistics are distributed as the order statistics of a sample of size J of the excess distribution. They consist of appropriate invokations of Talagrand's concentration inequality (Theorem 2.15). However, this theorem generally requires a uniform bound on the gradient of the relevant function. When Hill estimators are analysed as functions of independent exponential random variables, the partial derivatives depend on the points at which the von Mises function is evaluated. In order to get interesting bounds, it is worth conditioning on an intermediate order statistic.

Throughout this subsection, let ℓ be an integer larger than $\sqrt{\ln\log_2 n}$ and J an integer not larger than n. As we use the exponential representation of order statistics, besides Hill estimators, the random variables that appear in the main statements are order statistics of exponential samples, $Y_{(k)}$ will denote the kth order statistic of a standard exponential sample of size n (we agree on $Y_{(n+1)} = 0$).

Theorem 3.3, Propositions 3.9 and 3.10 complement each other in the following way. Theorem 3.3 is concerned with the supremum of the Hill process $\sqrt{i}\left|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]\right|$ for $\ell \le i \le k$. Note the use of random centering. The components of this process are shown to be sub-gamma using Talagrand's inequality, and then chaining is used to control the maximum of the process. Propositions 3.9 and 3.10 are concerned with conditional bias fluctuations, they state that the fluctuations of conditional expectations $|\mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}]|$ are small and even negligible with respect to the fluctuation of $\widehat{\gamma}(i)$.

The first theorem provides an exponential refinement of the variance bound stated in Proposition 3.1. However, as announced, there is a price to pay, statements hold conditionally on some order statistic.

In the sequel, let

$$\xi_n = c_1 \sqrt{\ln \log_2 n} + c_1',$$

where c_1 may be chosen not larger than 4 and c_1' not larger than 16.

Theorem 3.3. *Let T be a shorthand for $\exp(Y_{(J+1)})$. For some k, $\ell \le k \le J$, let*

$$Z = \max_{\ell \le i \le k} \sqrt{i} \left| \widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right] \right| .$$

Then, conditionally on T,

i) For $\ell \le i \le k$, the random variable

$$\sqrt{i} \left(\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T]\right)$$

is sub-gamma on both tails with variance factor $4\left(\gamma + 2\overline{\eta}\,(T)\right)^2$ and scale factor $\left(\gamma + \overline{\eta}\,(T)\right)/\sqrt{i}$.

ii) The random variable Z is sub-gamma on both tails with variance factor $4(\gamma + 2\overline{\eta}(T))^2$ and scale factor $(\gamma + 2\overline{\eta}(T))/\ell$ and

$$\mathbb{E}\left[Z \mid T\right] \le \xi_n \left(\gamma + 2\overline{\eta}(T)\right) . \tag{3.4}$$

Remark 3.5. If F is a pure Pareto distribution with shape parameter $\gamma > 0$, then $k\widehat{\gamma}(k)/\gamma$ is distributed according to a gamma distribution with shape parameter k and scale parameter 1. Tight and well-known tail bounds for gamma distributed random variables assert that

$$\mathbb{P}\left\{ |\widehat{\gamma}(k) - \mathbb{E}\left[\widehat{\gamma}(k)\right]| \ge \frac{\gamma}{\sqrt{k}} \left(\sqrt{2\ln\left(2/\delta\right)} + \frac{\ln\left(2/\delta\right)}{\sqrt{k}} \right) \right\} \le 2\delta\,.$$

Remark 3.6. If we choose $J = n$, all three statements hold unconditionally, but the variance factor may substantially exceed the upper bounds described in Proposition 3.1. Lemma 1 from [Drees and Kaufmann, 1998] reads as follows

$$\max_{2 \le i \le n} \sqrt{i} \left| \widehat{\gamma}(i) - \mathbb{E}\widehat{\gamma}(i) \right| = O_P\left(\sqrt{\ln \ln n}\right) .$$

The second and third statements in Theorem 3.3 provide a non-asymptotic counterpart to this lemma:

$$\mathbb{E}\left[\max_{2 \le i \le n} \sqrt{i} \left| \widehat{\gamma}(i) - \mathbb{E}\widehat{\gamma}(i) \right| \right] \le \left(c_1 \sqrt{2 \ln \log_2 n} + c_1' \right) \left(\gamma + 2\overline{\eta}(1)\right) ,$$

while the random variable in the expectation is sub-gamma.

Remark 3.7. Thanks to the Markov property, Statement i) reads as

$$\mathbb{P}\left\{ |Z - \mathbb{E}[Z \mid T]| \ge \left(\gamma + 2\overline{\eta}(T)\right) \left(\sqrt{8\ln\left(2/\delta\right)} + \frac{\ln\left(2/\delta\right)}{\ell} \right) \right\} \le \delta$$

where $0 < \delta < 1/2$. Combining Statements *ii)* and *iii)*, we also get

$$\mathbb{P}\left\{ Z \geq (\gamma + 2\overline{\eta}(T)) \left(\xi_n + \sqrt{8 \ln (2/\delta)} + \frac{\ln (2/\delta)}{\ell} \right) \right\} \leq \delta \,.$$

Remark 3.8. The reader may wonder whether resorting to the exponential representation and usual Chernoff bounding would not provide a simpler argument. The straightforward approach leads to the following conditional bound on the logarithmic moment generating function,

$$\begin{aligned} &\ln \mathbb{E}\left[\exp\left(\lambda \left(\widehat{\gamma}(k) - \mathbb{E}[\widehat{\gamma}(k) \mid Y_{(k+1)}] \right) \right) \mid Y_{(k+1)} \right] \\ &\leq \quad \frac{\left(\gamma + \overline{\eta}(\mathrm{e}^{Y_{(k+1)}})\right)^2}{2k\left(1 - \lambda(\gamma + \overline{\eta}(\mathrm{e}^{Y_{(k+1)}}))\right)} + \lambda\left(\overline{\eta}(\mathrm{e}^{Y_{(k+1)}}) - b\left(\mathrm{e}^{Y_{(k+1)}}\right) \right) \,. \end{aligned}$$

A similar statement holds for the lower tail. This leads to exponential bounds for deviation of the Hill estimator above $\mathbb{E}[\widehat{\gamma}(k) \mid Y_{(k+1)}] + \overline{\eta}(\mathrm{e}^{Y_{(k+1)}}) - b\left(\mathrm{e}^{Y_{(k+1)}}\right)$ that is, to control deviations of the Hill estimator above its expectation plus a term that may be of the order of magnitude of the bias.

Attempts to rewrite $\widehat{\gamma}(k) - \mathbb{E}[\widehat{\gamma}(k) \mid Y_{(k+1)}]$ as a sum of martingale increments $\mathbb{E}[\widehat{\gamma}(k) \mid Y_{(i)}] - \mathbb{E}[\widehat{\gamma}(k) \mid Y_{(i+1)}]$ for $1 \leq i \leq k$ and to exhibit an exponential supermartingale met the same impediments.

At the expense of inflating the variance factor, Theorem 2.15 provides a genuine (conditional) concentration inequality for Hill estimators. As we will deal with values of k for which bias exceeds the typical order of magnitudes of fluctations, this is relevant to our purpose.

The next propositions are concerned with the fluctuations of the conditional bias of Hill estimators. In both propositions, J satisfies $\ell \leq k \leq J \leq n$, and again $T = \exp\left(Y_{(J+1)}\right)$.

Proposition 3.9. *For all $1 \leq i \leq k$, conditionally on T,*

$$\mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}] - \mathbb{E}[\widehat{\gamma}(i)]$$

is sub-gamma on both tails with variance factor at most $16\overline{\eta}(T)^2/k$ and scale factor $2\overline{\eta}(T)/k$.

The last proposition deals with the maximum of centered conditional biases. The collection of centered conditional biases does not behave like partial sums.

Proposition 3.10. *Let the random variable Z be defined by*

$$Z = \max_{\ell \leq i \leq k} \left| \mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}] - \mathbb{E}[\widehat{\gamma}(i)] \right| \,.$$

Then,

i) Conditionally on T, Z is sub-gamma with variance factor $16\overline{\eta}(T)^2/k$ and scale factor $2\overline{\eta}(T)/k$.

ii)

$$\mathbb{E}Z \leq 4\sqrt{\frac{\mathbb{E}\left[\overline{\eta}(T)^2\right]}{k}}\,.$$

Remark 3.11. Statements *i)* and *ii)* in Proposition 3.10 can be summarized by the following inequality. For any $0 < \delta < 1/2$,

$$\mathbb{P}\left\{Z \geq \frac{4\overline{\eta}(1)}{\sqrt{k}}\left(1+\sqrt{2\ln(2/\delta)}+\frac{\ln(2/\delta)}{2\sqrt{k}}\right)\right\} \leq \delta\,. \tag{3.12}$$

Combining Theorem 3.3, Propositions 3.9 and 3.10 leads to another non-asymptotic perspective on Lemma 1 from [Drees and Kaufmann, 1998].

Corollary 3.13. *Let k be such that $\ell \leq k \leq n$ and let $T = \exp(Y_{(k+1)})$, then*

$$\mathbb{E}\left[\max_{\ell\leq i\leq k}\sqrt{i}\,|\widehat{\gamma}(i)-\mathbb{E}\widehat{\gamma}(i)|\right] \leq \xi_n\left(\gamma+2\mathbb{E}\overline{\eta}\,(T)\right)+4\sqrt{\mathbb{E}\left[\overline{\eta}\,(T)^2\right]}\,.$$

For $0 < \delta < 1/3$, with probability larger than $1-3\delta$,

$$\begin{aligned}
&\max_{\ell\leq i\leq k}\sqrt{i}\,|\widehat{\gamma}(i)-\mathbb{E}\widehat{\gamma}(i)| \\
&\leq \quad \xi_n(\gamma+2\overline{\eta}(n/k'))+4\sqrt{\mathbb{E}\left[\overline{\eta}(T)^2\right]} \\
&\quad +2\left(\gamma+2\overline{\eta}(n/k')+2\overline{\eta}(1)\right)\sqrt{2\ln(2/\delta)} \\
&\quad +\left(\frac{(\gamma+2\overline{\eta}(n/k'))}{\ell}+\frac{2\overline{\eta}(1)}{\sqrt{k}}\right)\ln(2/\delta)
\end{aligned}$$

where $k' = k + \frac{\ln(1/\delta)}{2}$.

3.3. Adaptive Hill estimation

We are now able to investigate the variant of the selection rule defined by (2.8) [Drees and Kaufmann, 1998] with $r_n = c_2\sqrt{\ln\ln n}$ where c_2 is a constant not smaller than $\sqrt{2}$.

The deterministic sequence of indices $(\widetilde{k}_n(r_n))$ is defined by:

$$\widetilde{k}_n(r_n) = \min\left\{k \in \{c_3\ln n,\ldots,n\}\colon\ |\mathbb{E}[\widehat{\gamma}(k)-\gamma]| > \frac{\gamma r_n}{\sqrt{k}}\right\}-1\,.$$

and the sequence $(\widetilde{k}_n(1))_n$ is defined by

$$\widetilde{k}_n(1) = \min\left\{k \in \{c_3\ln n,\ldots,n\}\colon\ |\mathbb{E}[\widehat{\gamma}(k)-\gamma]| > \frac{\gamma}{\sqrt{k}}\right\}-1\,.$$

Let $0 < \delta < 1/2$. The index $\widehat{k}_n$ is selected according to the following rule:

$$\widehat{k}_n = \min\left\{k \in \{c_3\ln n,\ldots,n\} \text{ and } \exists i \in \{\ell,\ldots,k\}\,, |\widehat{\gamma}(i)-\widehat{\gamma}(k)| > \frac{r_n(\delta)\widehat{\gamma}(i)}{\sqrt{i}}\right\}-1$$

where c_3 is a constant larger than 60 and $r_n(\delta) = 8\sqrt{2\ln\left((2/\delta)\log_2 n\right)}$. The tail index estimator is $\widehat{\gamma}(\widehat{k}_n)$.

Note that

$$\widetilde{k}_n(r_n) = \min\left\{k \in \{c_3 \ln n, \ldots, n\} \colon \max_{c_3 \ln n \le i \le k} \sqrt{i}\frac{|\mathbb{E}[\widehat{\gamma}(i) - \gamma]|}{\gamma} > r_n\right\} - 1\,.$$

while

$$\widetilde{k}_n(r_n) = \min\left\{k \in \{c_3 \ln n, \ldots, n\} \colon \max_{c_3 \ln n \le i \le k} \sqrt{i}\frac{|\mathbb{E}[\widehat{\gamma}(i) - \gamma]|}{\gamma} > 1\right\} - 1\,.$$

Thus, $\widehat{k}_n$ is an empirical version of $\widetilde{k}_n(r_n)$.

As tail adaptivity has a price (see Theorem 2.9), the ratio between the risk of the would-be adaptive estimator $\widehat{\gamma}(\widehat{k}_n)$ and the risk of $\widehat{\gamma}(\widetilde{k}_n(1))$ cannot be upper bounded by a constant factor, let alone by a factor close to 1. This is why in the next theorem, we compare the risk of $\widehat{\gamma}(\widehat{k}_n)$ with the risk of $\widehat{\gamma}(\widetilde{k}_n)$.

In the sequel, $\widetilde{k}_n$ stands for $\widetilde{k}_n(r_n)$. If the context is not clear, we specify $\widetilde{k}_n(1)$ or $\widetilde{k}_n(r_n)$. Recall that

$$\xi_n = c_1\sqrt{\ln\log_2 n} + c_1' \quad \text{with } c_1 \le 4 \text{ and } c_1' \le 16.$$

The next theorem describes a non-asymptotic risk bound for $\widehat{\gamma}(\widehat{k}_n)$.

Theorem 3.14. *Assume the sampling distribution $F \in \mathsf{MDA}(\gamma), \gamma > 0$ satisfies the von Mises condition with von Mises function η, and $\overline{\eta}(t) = \sup_{s \ge t}|\eta(s)|$.*

Assume that n is large enough so that

i) $\overline{\eta}(1) < \xi_n\gamma$,

ii) $\overline{\eta}\left(\frac{n}{\widetilde{k}_n + \frac{\ln(1/\delta)}{2}}\right) < \gamma/4$,

With probability larger than $1 - 3\delta$,

$$\left|\gamma - \widehat{\gamma}(\widehat{k}_n)\right| \le \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right|\left(1 + \frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\right) + \frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\gamma\,.$$

With probability larger than $1 - 4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n) - \gamma\right| \le \frac{2r_n(\delta)}{\sqrt{\widetilde{k}_n}}\gamma\left(1 + \alpha(\delta, n)\right)\,, \tag{3.15}$$

where

$$\begin{aligned}\alpha(\delta, n) \le\ & \frac{3}{16\sqrt{2}}\sqrt{\frac{\ln(2/\delta)}{\ln(2/\delta\log_2 n)}} + \frac{3}{16\sqrt{2c_3}}\frac{\ln(2/\delta)}{\sqrt{\ln n}\sqrt{\ln(2/\delta\log_2 n)}} \\ & + \frac{3}{2\sqrt{2c_3}}\sqrt{\frac{\ln(2/\delta)}{\ln n}} + \frac{3}{2c_3}\frac{\ln(2/\delta)}{\ln n}\,.\end{aligned}$$

Remark 3.16. For $0 < \delta < 1/2$,

$$\alpha(\delta, n) = o(1) \quad \text{as } n \to \infty \ .$$

Remark 3.17. If we assume that the bias b is ρ-regularly varying, then elaborating on Proposition 1 from [Drees and Kaufmann, 1998], the oracle index sequence $(k_n^*)_n$ and the sequence $(\widetilde{k}_n(1))_n$ are connected by

$$\lim_n \frac{\widetilde{k}_n(1)}{k_n^*} = (2|\rho|)^{1/(1+2|\rho|)}$$

and their quadratic risk are related by

$$\lim_n \frac{\mathbb{E}[(\gamma - \widehat{\gamma}(\widetilde{k}_n(1)))^2]}{\mathbb{E}[(\gamma - \widehat{\gamma}(k_n^*))^2]} = \frac{2}{2|\rho|+1}(2|\rho|)^{2|\rho|/(1+2|\rho|)} \ .$$

Thus if the bias is ρ-regularly varying, Theorem 3.14 provides us with a connection between the performance of the simple selection rule and the performance of the (asymptotically) optimal choice.

The next corollary upper bounds the risk of the preliminary estimator when we just have an upper bound on the bias.

Corollary 3.18. *Assume that for some $C > 0$ and $\rho < 0$, for all n, k,*

$$\left|\gamma - \mathbb{E}\widehat{\gamma}(k)\right| \le C\left(\frac{n}{k}\right)^{\rho} \ ,$$

then, there exists a constant $\kappa_{\delta,\rho}$ depending on δ and ρ such that, with probability larger than $1 - 4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n) - \gamma\right| \quad \le \quad \kappa_{\delta,\rho}\left(\frac{\gamma^2 \ln\left((2/\delta)\ln n\right)}{n}\right)^{|\rho|/(1+2|\rho|)} (1 + \alpha(\delta, n)) \quad .$$

Under the assumption that the bias of the Hill estimators is upper bounded by a power function, the performance of the data-driven estimator $\widehat{\gamma}(\widehat{k}_n)$ meets the information-theoretic lower bound of Theorem 2.9.

4. Proofs

4.1. Proof of Proposition 2.2

This proposition is a straightforward consequence of Rényi's representation of order statistics of standard exponential samples.

As F belongs to $\mathsf{MDA}(\gamma)$ and meets the von Mises condition, there exists a function η on $(1, \infty)$ with $\lim_{x\to\infty} \eta(x) = 0$ such that

$$U(x) = cx^{\gamma} \exp\left(\int_1^x \frac{\eta(s)}{s} \mathrm{d}s\right) \ ,$$

and

$$U(\mathrm{e}^y) = c \exp\left(\int_0^y (\gamma + \eta(\mathrm{e}^u)) \mathrm{d}u \right) .$$

Then,

$$\begin{aligned}
\widehat{\gamma}(k) &\stackrel{d}{=} \frac{1}{k} \sum_{i=1}^{k} i \frac{\log U\left(\mathrm{e}^{Y_{(i)}}\right)}{\log U\left(\mathrm{e}^{Y_{(i+1)}}\right)} \\
&\stackrel{d}{=} \frac{1}{k} \sum_{i=1}^{k} i \int_{Y_{(i+1)}}^{Y_{(i)}} (\gamma + \eta(\mathrm{e}^u)) \mathrm{d}u \\
&\stackrel{d}{=} \frac{1}{k} \sum_{i=1}^{k} i \int_{0}^{E_i/i} (\gamma + \eta(\mathrm{e}^{u+Y_{(i+1)}})) \mathrm{d}u \\
&\stackrel{d}{=} \frac{1}{k} \sum_{i=1}^{k} \int_{0}^{E_i} (\gamma + \eta(\mathrm{e}^{\frac{u}{i}+Y_{(i+1)}})) \mathrm{d}u \,.
\end{aligned}$$

4.2. Proof of Proposition 3.1

Let $Z = k\widehat{\gamma}(k)$. By the Pythagorean relation,

$$\mathrm{Var}(Z) = \mathbb{E}\left[\mathrm{Var}\left(Z \mid Y_{(k+1)}\right)\right] + \mathrm{Var}\left(\mathbb{E}[Z \mid Y_{(k+1)}]\right) .$$

Representation (2.4) asserts that, conditionally on $Y_{(k+1)}$, Z is distributed as a sum of independent, exponentially distributed random variables. Let E be an exponentially distributed random variable.

$$\begin{aligned}
&\mathrm{Var}\left(Z \mid Y_{(k+1)} = y\right) \\
&= k \,\mathrm{Var}\left(\gamma E + \int_0^E \eta(\mathrm{e}^{u+y}) \mathrm{d}u\right) \\
&= k\gamma^2 + 2k\gamma \,\mathrm{Cov}\left(E, \int_0^E \eta(\mathrm{e}^{u+y}) \mathrm{d}u\right) + \mathrm{Var}\left(\int_0^E \eta(\mathrm{e}^{u+y}) \mathrm{d}u\right) \\
&\leq k\gamma^2 + 2k\gamma\overline{\eta}(\mathrm{e}^y) + k\left(\overline{\eta}\left(\mathrm{e}^y\right)\right)^2 ,
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality and $\mathrm{Var}\big(\int_0^E \eta(\mathrm{e}^{y+u})\mathrm{d}u\big) \leq \overline{\eta}(\mathrm{e}^y)^2$. Taking expectation with respect to $Y_{(k+1)}$ leads to

$$\mathbb{E}\left[\mathrm{Var}\left(Z \mid Y_{(k+1)}\right)\right] \leq k\gamma^2 + 2k\gamma\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)\right] + k\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)^2\right] .$$

The last term in the Pythagorean decomposition is also handled using elementary arguments.

$$\mathbb{E}[Z \mid Y_{(k+1)}] - \mathbb{E}Z = k \int_0^\infty \mathrm{e}^{-u} \left(\eta\left(\mathrm{e}^{u+Y_{(k+1)}}\right) - \mathbb{E}\left[\eta\left(\mathrm{e}^{u+Y_{(k+1)}}\right)\right]\right) \mathrm{d}u \,.$$

As $Y_{(k+1)}$ is a function of independent exponential random variables ($Y_{(k+1)} = \sum_{i=k+1}^{n} E_i/i$), the variance of $\mathbb{E}[Z \mid Y_{(k+1)}]$ may be upper bounded using Poincaré inequality (Proposition 2.10)

$$\mathrm{Var}\left(\mathbb{E}[Z \mid Y_{(k+1)}]\right) \leq 4k\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)^2\right] .$$

In order to derive the lower bound, we first observe that

$$\mathrm{Var}(Z) \geq \mathbb{E}\left[\mathrm{Var}\left(Z \mid Y_{(k+1)}\right)\right] .$$

Now, using Cauchy-Schwarz inequality again

$$\begin{aligned}
\mathrm{Var}\left(Z \mid Y_{(k+1)} = y\right) &\geq k\gamma^2 - 2k\gamma\left|\mathrm{Cov}\big(E, \int_0^E \eta(\mathrm{e}^{u+y})\mathrm{d}u\big)\right| \\
&\geq k\gamma^2 - 2k\gamma\left(\mathrm{Var}\left(\int_0^E \eta(\mathrm{e}^{u+y})\mathrm{d}u\right)\right)^{1/2} \\
&\geq k\gamma^2 - 2k\gamma\overline{\eta}(\mathrm{e}^{y}) .
\end{aligned}$$

4.3. Proof of Theorem 3.3

In the proofs of Theorem 3.3 and Propositon 3.10, we will use the next maximal inequality.

Proposition 4.1. *[Boucheron et al., 2013, Corollary 2.6] Let $Z_1, \ldots, Z_N$ be real-valued random variables belonging to $\Gamma_+(v, c)$. Then*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} Z_i\right] \leq \sqrt{2v \log N} + c \log N .$$

Proofs follow a common pattern. In order to check that some random variable is sub-gamma, we rely on its representation as a function of independent exponential variables and compute partial derivatives, derive convenient upper bounds on the squared Euclidean norm and the supremum norm of the gradient and then invoke Theorem 2.15.

At some point, we will use the next corollary of Theorem 2.15.

Corollary 4.2. *If f is an almost everywhere differentiable function on $\mathbb{R}$ with uniformly bounded derivative f', then $f(Y_{(k+1)})$ is sub-gamma with variance factor $4\|f'\|_\infty^2/k$ and scale factor $\|f'\|_\infty/k$.*

Proof of Theorem 3.3. We start from the exponential representation of Hill estimators (Proposition 2.2) and represent all $\widehat{\gamma}(i)$ as functions of independent random variables $E_1, \ldots, E_k, Y_{(k+1)}$ where the $E_j, 1 \leq j \leq k$ are standard exponentially distributed and $Y_{(k+1)}$ is distributed like the $(k+1)$th largest order statistic of an n-sample of the standard exponential distribution.

$$
\begin{aligned}
i\widehat{\gamma}(i) &= \sum_{j=1}^{i} \int_0^{E_j} \left(\gamma + \eta(\mathrm{e}^{\frac{u}{j}+Y_{(j+1)}})\right) \mathrm{d}u \\
&= \sum_{j=1}^{i} \int_0^{E_j} \left(\gamma + \eta(\mathrm{e}^{\frac{u}{j}+\sum_{m=j+1}^{k} \frac{E_m}{m}+Y_{(k+1)}})\right) \mathrm{d}u\,.
\end{aligned}
$$

Let i' be such that $0 \le i' < i$, agree on $\widehat{\gamma}(0) = 0$. Then,

$$
i\widehat{\gamma}(i) - i'\widehat{\gamma}(i') = \sum_{j=i'+1}^{i} \int_0^{E_j} (\gamma + \eta(\mathrm{e}^{\frac{u}{j}+\sum_{m=j+1}^{k} \frac{E_m}{m}+Y_{(k+1)}}))\mathrm{d}u\,.
$$

Letting

$$
g(x_{i'+1}, \ldots, x_k) = \sum_{j=i'+1}^{i} \int_0^{x_j} (\gamma + \eta(\mathrm{e}^{\frac{u}{j}+\sum_{m=j+1}^{k} \frac{x_m}{m}+Y_{(k+1)}}))\mathrm{d}u\,,
$$

a few lines of computations lead to

$$
\partial_j g = \begin{cases} 0 & \text{for } j \le i' \\ \gamma + \frac{1}{j}\sum_{m=i'+1}^{j} \eta(\mathrm{e}^{Y_{(m)}}) & \text{for } i' < j \le i \\ \frac{1}{j}\sum_{m=i'+1}^{i} \eta(\mathrm{e}^{Y_{(m)}}) & \text{for } i < j \le k \end{cases}
$$

which entails that

$$
|\partial_j g| \le \begin{cases} \gamma + \overline{\eta}(\mathrm{e}^{Y_{(J+1)}}) & \text{for } i' < j \le i \\ \frac{i-i'}{j}\overline{\eta}(\mathrm{e}^{Y_{(J+1)}}) & \text{for } i < j \le k\,. \end{cases}
$$

Recalling that $T = \exp\left(Y_{(J+1)}\right)$, this can be summarised by

$$
\left|\frac{\partial g}{\partial x_j}\right| \le \gamma + \overline{\eta}(T)
$$

and

$$
\sum_{j=1}^{k} \left|\frac{\partial g}{\partial x_j}\right|^2 \le (i - i')\left(\gamma + 2\overline{\eta}(T)\right)^2\,.
$$

Theorem 2.15 now allows us to establish that, conditionally on T, the centered version of $i\widehat{\gamma}(i) - i'\widehat{\gamma}(i')$ is sub-gamma on both tails with variance factor $4|i - i'|(\gamma + 2\overline{\eta}(T))^2$ and scale factor $(\gamma + \overline{\eta}(T))$. Using Theorem 2.15 conditionally on T, and choosing $i' = 0$,

$$
\mathbb{P}\left\{|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid T\right]| \ge \frac{\gamma + 2\overline{\eta}\,(T)}{\sqrt{i}}\left(\sqrt{8s} + \frac{s}{\sqrt{i}}\right) \mid T\right\} \le 2\mathrm{e}^{-s}\,.
$$

Taking expectation on both sides, this implies that

$$\mathbb{P}\left\{|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid T\right]| \geq \frac{\gamma + 2\overline{\eta}\,(T)}{\sqrt{i}}\left(\sqrt{8s} + \frac{s}{\sqrt{i}}\right)\right\} \leq 2\mathrm{e}^{-s}\,.$$

The proof of the upper bound on $\mathbb{E}[Z \mid T]$ in Statement ii) from Theorem 3.3 relies on standard chaining techniques from the theory of empirical processes and uses repeatedly the concentration Theorem 2.15 for smooth functions of independent exponential random variables and the maximal inequality for sub-gamma random variables (Proposition 4.1).

Recall that

$$Z = \max_{\ell \leq i \leq k} \sqrt{i}\left|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]\right|\,.$$

As it is commonplace in the analysis of normalized empirical processes [See Giné and Koltchinskii, 2006, Massart, 2007, van de Geer, 2000, and references therein], we peel the index set over which the maximum is computed.

Let $\mathcal{L}_n = \{\lfloor \log_2(\ell)\rfloor, \ldots, \lfloor \log_2(k)\rfloor\}$. For all $j \in \mathcal{L}_n$, let $\mathcal{S}_j = \{\ell \vee 2^j, \ldots, k \wedge 2^{j+1} - 1\}$ and define Z_j as

$$Z_j = \max_{i \in \mathcal{S}_j} \sqrt{i}\left|\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]\right|\,.$$

Then,

$$\begin{aligned}\mathbb{E}[Z \mid Y_{(k+1)}] &= \mathbb{E}[\max_{j \in \mathcal{L}_n} Z_j \mid Y_{(k+1)}]] \\ &\leq \mathbb{E}[\max_{j \in \mathcal{L}_n}(Z_j - \mathbb{E}[Z_j \mid Y_{(k+1)}]) \mid Y_{(k+1)}]] + \max_{j \in \mathcal{L}_n} \mathbb{E}[Z_j \mid Y_{(k+1)}]]\,.\end{aligned}$$

We now derive upper bounds on both summands by resorting to the maximum inequality for sub-gamma random variables (Proposition 4.1). We first bound $\mathbb{E}[Z_j \mid Y_{(k+1)}]$ for $j \in \mathcal{L}_n$.

Fix $j \in \mathcal{L}_n$,

$$\max_{i \in \mathcal{S}_j} \sqrt{i}\left|\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}]\right| \leq \frac{1}{2^{j/2}} \max_{i \in \mathcal{S}_j} i\left|\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}]\right|\,.$$

In order to alleviate notation, let $W(i) = i\left(\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}]\right)$ for $i \in \mathcal{S}_j$. For $i \in \mathcal{S}_j$, let

$$i = 2^j + \sum_{m=1}^{i} b_m 2^{j-m} \quad \text{where } b_m \in \{0,1\}$$

be the binary expansion of i. Then, for $h \in \{0, \ldots, j\}$, let $\pi_h(i)$ be defined by

$$\pi_h(i) = 2^j + \sum_{m=1}^{h} b_m 2^{j-m}$$

so that $\pi_j(i) = i$, $\pi_0(i) = 2^j$ and $0 \leq \pi_{h+1}(i) - \pi_h(i) \leq 2^{j-h-1}$.

Using that $W(\pi_0(i))$ does not depend on i and that $\mathbb{E}\left[W(\pi_0(i)) \mid Y_{(k+1)}\right] = 0$,

$$\begin{aligned}
&\mathbb{E}\left[\max_{i\in\mathcal{S}_j} i\left(\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]\right) \mid Y_{(k+1)}\right] \\
&\quad= \mathbb{E}\left[\max_{i\in\mathcal{S}_j} W(i) \mid Y_{(k+1)}\right] \\
&\quad= \mathbb{E}\left[\max_{i\in\mathcal{S}_j} W(\pi_j(i)) - W(\pi_0(i)) \mid Y_{(k+1)}\right] \\
&\quad= \mathbb{E}\left[\max_{i\in\mathcal{S}_j} \sum_{h=0}^{j-1} (W(\pi_{h+1}(i)) - W(\pi_h(i))) \mid Y_{(k+1)}\right] \\
&\quad\leq \sum_{h=0}^{j-1} \mathbb{E}\left[\max_{i\in\mathcal{S}_j} (W(\pi_{h+1}(i)) - W(\pi_h(i))) \mid Y_{(k+1)}\right] .
\end{aligned}$$

Now for each $h \in \{0, \ldots, j-1\}$, the maximum is taken over 2^h random variables which are sub-gamma with variance factor $4 \times 2^{j-h-1}(\gamma + 2\overline{\eta}(T))^2$ and scale factor $(\gamma + \overline{\eta}(T))$. By Proposition 4.1,

$$\begin{aligned}
&\mathbb{E}\left[\max_{i\in\mathcal{S}_j} i\left(\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]\right) \mid Y_{(k+1)}\right] \\
&\quad\leq \sum_{h=0}^{j-1} \left((\gamma + 2\overline{\eta}(T))\sqrt{8h2^{(j-h-1)}\ln 2} + (\gamma + \overline{\eta}(T))h\ln 2\right) \\
&\quad\leq 8\,(\gamma + 2\overline{\eta}(T))2^{j/2} .
\end{aligned}$$

So that for all $j \in \mathcal{L}_n$,

$$\mathbb{E}\left[\max_{i\in\mathcal{S}_j} \sqrt{i}(\widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]) \mid Y_{(k+1)}\right] \leq 8(\gamma + 2\overline{\eta}(T))$$

and

$$\mathbb{E}[Z_j \mid Y_{(k+1)}] \leq 16(\gamma + 2\overline{\eta}(T)) .$$

In order to prove Statement iii), we check that for each $j \in \mathcal{L}_n$, Z_j is sub-gamma on the right-tail with variance factor at most $4\,(\gamma + 2\overline{\eta}(T))^2$ and scale factor not larger than $(\gamma + 2\overline{\eta}(T))/\ell$. Under the von Mises Condition (2.1), the sampling distribution is absolutely continuous with respect to Lebesgue measure. For almost every sample, the maximum defining Z_j is attained at a single index $i \in \mathcal{S}_j$. Starting again from the exponential representation, and repeating the computation of partial derivatives, we obtain the desired bounds.

By Proposition 4.1,

$$\begin{aligned}
\mathbb{E}\left[\max_{j\in\mathcal{L}_n} (Z_j - \mathbb{E}[Z_j \mid Y_{(k+1)}]) \mid Y_{(k+1)}\right] &\leq \left(\sqrt{8\ln|\mathcal{L}_n|} + \frac{\ln|\mathcal{L}_n|}{\ell}\right)(\gamma + 2\overline{\eta}(T)) \\
&\leq 6(\sqrt{\ln|\mathcal{L}_n|}\,(\gamma + 2\overline{\eta}(T)) .
\end{aligned}$$

Combining the different bounds leads to Inequality (3.4). □

4.4. Proof of Proposition 3.9

Adopting again the exponential representation, $\mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]$ is an $Y_{(k+1)}$-measurable random variable, say $f(Y_{(k+1)})$.

As a function f of $Y_{(k+1)} = y$, the conditional expectation $\mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right]$ reads as

$$f(y) \quad = \quad \frac{1}{i}\sum_{j=1}^{i} \int_{[0,\infty)} \cdots \int \mathrm{e}^{-\sum_{m=j}^{k} u_m} \eta\left(\mathrm{e}^{y+\sum_{m=j}^{k} \frac{u_m}{m}}\right) \mathrm{d}u_j \ldots \mathrm{d}u_k \,.$$

Its derivative with respect to y is readily computed, and after integration by parts and handling a telescoping sum, it reads as

$$\begin{aligned} & f'(y) \\ & = \quad \frac{1}{i}\sum_{j=1}^{i} \int_{[0,\infty)} \cdots \int \mathrm{e}^{-\sum_{m=j}^{k} u_m} \left(\eta\left(\mathrm{e}^{y+\sum_{m=j}^{k} \frac{u_m}{m}}\right)\right. \\ & \qquad\qquad \left. -\eta\left(\mathrm{e}^{y+\sum_{m=j+1}^{k} \frac{u_m}{m}}\right)\right) \mathrm{d}u_j \ldots \mathrm{d}u_k \,. \end{aligned}$$

A conservative upper-bound on $|f'(y)|$ is $2\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})$ which is upper bounded by $2\overline{\eta}(T)$. The statement of the proposition then follows from Proposition 4.2.

A byproduct of the proof is the next variance bound, for $i \leq k$,

$$\mathrm{Var}\left(\mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}] \mid T\right) \leq \frac{16}{k}\mathbb{E}\left[\overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)^2 \mid T\right] \,.$$

4.5. Proof of Proposition 3.10

In the proof, Δ_i denotes the spacing $Y_{(i)} - Y_{(k+1)}$, $\mathbb{E}_{\Delta_i}$ the expectation with respect to Δ_i, $Y'_{(k+1)}$ an independent copy of $Y_{(k+1)}$, and $\mathbb{E}'$ the expectation with respect to $Y'_{(k+1)}$. We will also use the next lemma.

Lemma 4.3. *Let X be a non-negative random variable, and $a, b \in [0,\infty]$, then*

$$\mathbb{E}\left[\mathrm{e}^{X}\left|\int_{\mathrm{e}^{X+a}}^{\mathrm{e}^{X+b}} \frac{\eta(v)}{v^2}\mathrm{d}v\right|\right] \leq \overline{\eta}\left(\mathrm{e}^{a\wedge b}\right)\left|\mathrm{e}^{-a} - \mathrm{e}^{-b}\right| \,.$$

Let

$$Z = \max_{\ell \leq i \leq k} \left|\mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}] - \mathbb{E}[\widehat{\gamma}(i)]\right|$$

and recall that $b(t) = t\int_t^\infty \eta(v)/v^2 \mathrm{d}v$. Then

$$\begin{aligned}
&\left|\mathbb{E}[\widehat{\gamma}(i) \mid Y_{(k+1)}] - \mathbb{E}[\widehat{\gamma}(i)]\right| \\
&\quad = \left|\mathbb{E}[b\left(\mathrm{e}^{Y_{(i+1)}}\right) \mid Y_{(k+1)}] - \mathbb{E}[b\left(\mathrm{e}^{Y_{(i+1)}}\right)]\right| \\
&\quad = \left|\mathrm{e}^{Y_{(k+1)}}\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^\infty \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right. \\
&\qquad \left. -\mathbb{E}'\left[\mathrm{e}^{Y'_{(k+1)}}\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y'_{(k+1)}}}^\infty \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right]\right| \\
&\quad \le \left|\mathbb{E}'\left[\left(\mathrm{e}^{Y_{(k+1)}} - \mathrm{e}^{Y'_{(k+1)}}\right)\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^\infty \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right]\right| \\
&\qquad + \left|\mathbb{E}'\left[\mathrm{e}^{Y'_{(k+1)}}\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^{\mathrm{e}^{\Delta_i+Y'_{(k+1)}}} \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right]\right| \ .
\end{aligned}$$

The expectation of Z is thus upper bounded by the following sum

$$\begin{aligned}
\mathbb{E}Z \ \le \ & \underbrace{\mathbb{E}\left[\max_{\ell\le i\le k}\left|\mathbb{E}'\left[\left(\mathrm{e}^{Y_{(k+1)}} - \mathrm{e}^{Y'_{(k+1)}}\right)\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^\infty \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right]\right|\right]}_{:=(\mathrm{I})} \\
& + \underbrace{\mathbb{E}\left[\max_{\ell\le i\le k}\left|\mathbb{E}'\left[\mathrm{e}^{Y'_{(k+1)}}\mathbb{E}_{\Delta_i}\left[\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^{\mathrm{e}^{\Delta_i+Y'_{(k+1)}}} \frac{\eta(v)}{v^2}\mathrm{d}v\right]\right]\right|\right]}_{:=(\mathrm{II})} \ .
\end{aligned}$$

In the sequel, we use twice the next upper bound

$$\mathbb{E}\left[\left(1 - \mathrm{e}^{Y'_{(k+1)}-Y_{(k+1)}}\right)^2\right] \le \frac{3}{k} \ . \tag{4.4}$$

By Lemma 4.3,

$$\begin{aligned}
(\mathrm{I}) \ &\le \ \mathbb{E}\left[\max_{\ell\le i\le k}\mathbb{E}'\left[\left|\mathrm{e}^{Y_{(k+1)}} - \mathrm{e}^{Y'_{(k+1)}}\right|\mathbb{E}_{\Delta_i}\left[\left|\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^\infty \frac{\eta(v)}{v^2}\mathrm{d}v\right|\right]\right]\right] \\
&\le \ \mathbb{E}\left[\max_{\ell\le i\le k}\mathbb{E}'\left[\left|\mathrm{e}^{Y_{(k+1)}} - \mathrm{e}^{Y'_{(k+1)}}\right| \times \overline{\eta}\left(\mathrm{e}^{Y_{(k+1)}}\right)\mathrm{e}^{-Y_{(k+1)}}\right]\right] \\
&= \ \mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})\left|1 - \mathrm{e}^{Y'_{(k+1)}-Y_{(k+1)}}\right|\right] \\
&\le \ \sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2\right]}\sqrt{\mathbb{E}\left[\left|1 - \mathrm{e}^{Y'_{(k+1)}-Y_{(k+1)}}\right|^2\right]}
\end{aligned}$$

and

$$\begin{aligned}
(\text{II}) &\leq \mathbb{E}\left[\max_{\ell\leq i\leq k}\mathbb{E}'\left[\mathrm{e}^{Y'_{(k+1)}}\mathbb{E}_{\Delta_i}\left[\left|\mathrm{e}^{\Delta_i}\int_{\mathrm{e}^{\Delta_i+Y_{(k+1)}}}^{\mathrm{e}^{\Delta_i+Y'_{(k+1)}}}\frac{\eta(v)}{v^2}\mathrm{d}v\right|\right]\right]\right] \\
&\leq \mathbb{E}\left[\mathrm{e}^{Y'_{(k+1)}}\overline{\eta}(\mathrm{e}^{Y_{(k+1)}\wedge Y'_{(k+1)}})\left|\mathrm{e}^{-Y_{(k+1)}}-\mathrm{e}^{-Y'_{(k+1)}}\right|\right] \\
&= \mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}\wedge Y'_{(k+1)}})\left|\mathrm{e}^{Y'_{(k+1)}-Y_{(k+1)}}-1\right|\right] \\
&\leq \sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2\right]}\sqrt{\mathbb{E}\left[\left|1-\mathrm{e}^{Y'_{(k+1)}-Y_{(k+1)}}\right|^2\right]}.
\end{aligned}$$

Finally, thanks to Inequality (4.4),

$$\mathbb{E}Z \leq \frac{2C\sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}\wedge Y'_{(k+1)}})^2\right]}}{\sqrt{k}}$$

where the constant C can be chosen not larger than 3.

4.6. Proof of Corollary 3.13

We first check that, with probability larger than $1-\delta$, $Y_{(k+1)} \geq n/k'$ where $k' = k + \ln(\delta)/2$. Starting from the proof of Proposition 4.3 from [Boucheron and Thomas, 2012] and recalling that $\operatorname{arsinh}(x) = \ln(x+\sqrt{1+x^2})$, a few lines of computation yields that, with probability larger than $1-\delta$, for $\ln 2 \leq z \leq \ln\left(\frac{n}{k}\right)$,

$$\mathbb{P}\left\{\exp\left(Y_{(k+1)}\right) \geq \frac{n}{k}\mathrm{e}^{-z}\right\} \geq 1-\exp\left(-\frac{k(\mathrm{e}^z-1)^2}{2\mathrm{e}^z}\right).$$

Solving

$$\delta = \exp\left(-\frac{k(\mathrm{e}^z-1)^2}{2\mathrm{e}^z}\right)$$

leads to

$$z = 2\operatorname{arsinh}\left(\sqrt{\frac{\ln(1/\delta)}{2k}}\right) = 2\ln\left(\sqrt{\frac{\ln(1/\delta)}{2k}}+\sqrt{1+\frac{\ln(1/\delta)}{2k}}\right),$$

so that

$$\mathrm{e}^z \leq 1+\frac{\ln(1/\delta)}{2k}.$$

Now,

$$\begin{aligned}
&\max_{\ell\leq i\leq k}\sqrt{i}\left|\widehat{\gamma}(i)-\mathbb{E}\widehat{\gamma}(i)\right| \\
&\quad\leq \max_{\ell\leq i\leq k}\sqrt{i}\left|\widehat{\gamma}(i)-\mathbb{E}[\widehat{\gamma}(i)\mid Y_{(k+1)}]\right| + \max_{\ell\leq i\leq k}\sqrt{i}\left|\mathbb{E}[\widehat{\gamma}(i)]-\mathbb{E}[\widehat{\gamma}(i)\mid Y_{(k+1)}]\right|.
\end{aligned}$$

By Theorem 3.3, with probability $1-2\delta$,

$$\max_{\ell\le i\le k}\sqrt{i}\left|\widehat{\gamma}(i)-\mathbb{E}[\widehat{\gamma}(i)\mid Y_{(k+1)}]\right|\le(\gamma+2\overline{\eta}(n/k'))\left(\xi_n+\sqrt{8\ln(2/\delta)}+\frac{\ln(2/\delta)}{\ell}\right).$$

Then, by Proposition 3.10, with probability larger than $1-\delta$,

$$\max_{\ell\le i\le k}\sqrt{i}\left|\mathbb{E}[\widehat{\gamma}(i)]-\mathbb{E}[\widehat{\gamma}(i)\mid Y_{(k+1)}]\right|\le 4\sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2\right]}+2\overline{\eta}(1)\left(2\sqrt{2\ln(2/\delta)}+\frac{\ln(2/\delta)}{\sqrt{k}}\right).$$

Combining all tail bounds leads to

$$\begin{aligned}
&\max_{\ell\le i\le k}\sqrt{i}\left|\widehat{\gamma}(i)-\mathbb{E}\widehat{\gamma}(i)\right|\\
&\le\ (\gamma+2\overline{\eta}(n/k'))\left(\xi_n+\sqrt{8\ln(2/\delta)}+\frac{\ln(2/\delta)}{\ell}\right)\\
&\quad+4\sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2\right]}++2\overline{\eta}(1)\left(2\sqrt{2\ln(2/\delta)}+\frac{\ln(2/\delta)}{\sqrt{k}}\right)\\
&\le\ \xi_n(\gamma+2\overline{\eta}(n/k'))+4\sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2\right]}+2(\gamma+2\overline{\eta}(n/k')+2\overline{\eta}(1))\sqrt{2\ln(2/\delta)}\\
&\quad+\left(\frac{(\gamma+2\overline{\eta}(n/k'))}{\ell}+\frac{\overline{\eta}(1)}{\sqrt{k}}\right)\ln(2/\delta)
\end{aligned}$$

with probability larger than $1-3\delta$.

4.7. *Proof of Theorem 3.14*

Throughout this proof, let

$$\begin{aligned}
T_n&=\exp\left(Y_{(\widetilde{k}_n+1)}\right)\\
\xi_n&=c_1\sqrt{\ln\log_2 n}+c_1'\quad\text{with } c_1\le 4 \text{ and } c_1'\le 16,\\
z_\delta&=\xi_n+\sqrt{8\ln(2/\delta)}+\frac{\ln(2/\delta)}{c_3\ln n}\\
y_\delta&=1+\sqrt{2\ln(2/\delta)}+\frac{\ln(2/\delta)}{2\sqrt{c_3\ln n}}.
\end{aligned}$$

Let us consider the event $E_1\cap E_2\cap E_3$ as

$$\begin{aligned}
E_1&=\left\{c_3\ln n\le i\le\widetilde{k}_n,\ \sqrt{i}\left|\widehat{\gamma}(i)-\mathbb{E}[\widehat{\gamma}(i)\mid T_n]\right|\le(\gamma+2\overline{\eta}(T_n))\,z_\delta\right\},\\
E_2&=\left\{c_3\ln n\le i\le k,\ \sqrt{k}\left|\mathbb{E}[\widehat{\gamma}(i)-\widehat{\gamma}(k)\mid T_n]-\mathbb{E}[\widehat{\gamma}(i)-\widehat{\gamma}(k)]\right|\le 8\overline{\eta}(1)y_\delta\right\},\\
E_3&=\left\{T_n\ge n/\left(\widetilde{k}_n+\ln(1/\delta)/2\right)\right\}.
\end{aligned}$$

$E_1\cap E_2\cap E_3$ has probability at least $1-3\delta$. Indeed, by Theorem 3.3, $\mathbb{P}(E_1)\ge 1-\delta$, by Theorem 3.10, $\mathbb{P}(E_2)\ge 1-\delta$ and thanks to the beginning of the proof of Corollary 3.13, $\mathbb{P}(E_3)\ge 1-\delta$.

Clearly, $E_1 = E_1(\delta, n)$, $E_2 = E_2(\delta, n)$, and $E_3 = E_3(\delta, n)$. However, since δ and n are fixed, we simply denote them E_1, E_2 and E_3 to alleviate the notations.

We check that under $E_1 \cap E_2 \cap E_3$, the selected index is not smaller than $\widetilde{k}_n$. This amounts to check that for all $k \leq \widetilde{k}_n - 1$, and for all $i \in \{\lceil c_3 \ln n \rceil, \ldots, k\}$,

$$\sqrt{i}\,|\widehat{\gamma}(i) - \widehat{\gamma}(k)| \leq r_n(\delta)\widehat{\gamma}(i)\,.$$

Now, under $E_1 \cap E_3$, $\overline{\eta}(T_n) \leq \gamma/4$ and conditionally on T_n, the process $(|\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T_n]|)_i$ can be controlled: for all $i \in \{\lceil c_3 \ln n \rceil, \ldots, k\}$

$$|\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T_n]| \;\leq \frac{(\gamma + 2\overline{\eta}(T_n))}{\sqrt{i}} z_\delta \leq \frac{3\gamma}{2\sqrt{i}} z_\delta \;. \tag{4.5}$$

For all $i \in \{\lceil c_3 \ln n \rceil, \ldots, \widetilde{k}_n\}$,

$$\begin{aligned}
\gamma - \widehat{\gamma}(i) &\leq |\gamma - \mathbb{E}[\widehat{\gamma}(i) \mid T_n]| + |\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T_n]| \\
&\leq \overline{\eta}(T_n) + \frac{3\gamma}{2\sqrt{i}} z_\delta \\
&\leq \frac{\gamma}{4} + \frac{3\gamma}{2\sqrt{i}} z_\delta
\end{aligned}$$

so that

$$\frac{\widehat{\gamma}(i)}{\gamma} \geq \frac{3}{4}\left(1 - \frac{2z_\delta}{\sqrt{i}}\right) .$$

Meanwhile, for all $k \leq \widetilde{k}_n - 1$ and for all $i \in \{\lceil c_3 \ln n \rceil, \ldots, k\}$,

$$\begin{aligned}
&|\widehat{\gamma}(i) - \widehat{\gamma}(k)| \\
&\leq \underbrace{\left|\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T_n]\right|}_{\text{(I)}} + \underbrace{|\mathbb{E}[\widehat{\gamma}(i) - \widehat{\gamma}(k) \mid T_n]|}_{\text{(II)}} + \underbrace{|\widehat{\gamma}(k) - \mathbb{E}[\widehat{\gamma}(k) \mid T_n]|}_{\text{(III)}} .
\end{aligned}$$

Using again (4.5), under $E_1 \cap E_3$,

$$\text{(I)} + \text{(III)} \leq \frac{3\gamma}{2} z_\delta \left(\frac{1}{\sqrt{i}} + \frac{1}{\sqrt{k}}\right) \leq \frac{3\gamma}{\sqrt{i}} z_\delta \;.$$

Now, under E_2, thanks to Assumption i) in the theorem statement,

$$\begin{aligned}
&\text{(II)} \\
&\quad\leq |\mathbb{E}[\widehat{\gamma}(i) - \widehat{\gamma}(k) \mid T_n] - \mathbb{E}[\widehat{\gamma}(i) - \widehat{\gamma}(k)]| + |\mathbb{E}[\widehat{\gamma}(i) - \gamma]| + |\mathbb{E}[\gamma - \widehat{\gamma}(k)]| \\
&\quad\leq \frac{8\overline{\eta}(1)}{\sqrt{k}} y_\delta + \gamma r_n \left(\frac{1}{\sqrt{i}} + \frac{1}{\sqrt{k}}\right) \\
&\quad\leq \frac{8\xi_n \gamma}{\sqrt{k}} y_\delta + \frac{2r_n\gamma}{\sqrt{i}} .
\end{aligned}$$

Plugging upper bounds on (I), (II) and (III), it comes that under E_δ, for all $k \leq \widetilde{k}_n - 1$ and for all $i \in \{\lceil c_3 \ln n \rceil, \ldots, k\}$,

$$\begin{aligned}
\sqrt{i}\frac{|\widehat{\gamma}(i) - \widehat{\gamma}(k)|}{\gamma} &\leq 3z_\delta + 8\xi_n y_\delta + 2r_n \\
&\leq 3\left(\xi_n + \sqrt{8\ln(2/\delta)} + \frac{\ln(2/\delta)}{c_3 \ln n}\right) + 8\xi_n y_\delta + 2r_n \\
&\leq \xi_n(3 + 8y_\delta) + 2r_n + 3\sqrt{8\ln(2/\delta)} + \frac{3\ln(2/\delta)}{c_3 \ln n}\,.
\end{aligned}$$

In order to warrant that under $E_1 \cap E_2 \cap E_3$, for all $k \leq \widetilde{k}_n - 1$ and for all i such that $c_3 \ln n \leq i \leq k$, $\sqrt{i}\,|\widehat{\gamma}(i) - \widehat{\gamma}(k)| \leq r_n(\delta)\widehat{\gamma}(i)$, it is enough to have

$$r_n(\delta)\left(1 - \frac{2z_\delta}{\sqrt{c_3 \ln n}}\right) \leq \frac{4}{3}\left(\xi_n(3 + 8y_\delta) + 2r_n + 3\sqrt{8\ln(2/\delta)} + \frac{3\ln(2/\delta)}{c_3 \ln n}\right)$$

which holds by definition of $r_n(\delta)$.

We now check that the risk of $\widehat{\gamma}(\widehat{k}_n)$ is not much larger than the risk of $\widehat{\gamma}(\widetilde{k}_n)$.

$$\begin{aligned}
\left|\gamma - \widehat{\gamma}(\widehat{k}_n)\right| &\leq \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right| + \left|\widehat{\gamma}(\widetilde{k}_n) - \widehat{\gamma}(\widehat{k}_n)\right| \\
&\leq \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right| + \frac{r_n(\delta)\widehat{\gamma}(\widetilde{k}_n)}{\sqrt{\widetilde{k}_n}} \\
&\leq \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right|\left(1 + \frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\right) + \frac{r_n(\delta)\gamma}{\sqrt{\widetilde{k}_n}}\,.
\end{aligned}$$

Therefore, with probability larger than $1 - 3\delta$,

$$\left|\gamma - \widehat{\gamma}(\widehat{k}_n)\right| \leq \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right|\left(1 + \frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\right) + \frac{r_n(\delta)\gamma}{\sqrt{\widetilde{k}_n}}\,. \tag{4.6}$$

Now, consider the event $E_1 \cap E_2 \cap E_3 \cap E_4$ with E_1, E_2 and, E_3 defined as in the beginning of the proof and

$$E_4 = \left\{\sqrt{\widetilde{k}_n}\left|\widehat{\gamma}(\widetilde{k}_n) - \mathbb{E}[\widehat{\gamma}(\widetilde{k}_n) \mid T_n]\right| \leq (\gamma + 2\overline{\eta}(T_n))\left(\sqrt{8\ln(2/\delta)} + \frac{\ln(2/\delta)}{\sqrt{\widetilde{k}_n}}\right)\right\}\,.$$

Since, $\mathbb{P}(E_4) \geq 1 - \delta$ thanks to Statement i) from Theorem 3.3, the event $E_1 \cap E_2 \cap E_3 \cap E_4$ has probability at least $1 - 4\delta$.

Then, by definition of $\widetilde{k}_n$, $|\gamma - \mathbb{E}\widehat{\gamma}(\tilde{k}_n)| \leq \gamma r_n/\sqrt{\widetilde{k}_n}$, under $E_3 \cap E_4$,

$$\begin{aligned}
\left|\widehat{\gamma}(\tilde{k}_n) - \gamma\right| &\leq |\gamma - \mathbb{E}\widehat{\gamma}(\tilde{k}_n)| + |\widehat{\gamma}(\tilde{k}_n) - \mathbb{E}\widehat{\gamma}(\tilde{k}_n)| \\
&\leq \frac{1}{\sqrt{\widetilde{k}_n}}\left(\gamma r_n + (\gamma + 2\overline{\eta}(T_n))\left(\sqrt{8\ln(2/\delta)} + \frac{\ln(2/\delta)}{\sqrt{\widetilde{k}_n}}\right)\right) \\
&\leq \frac{\gamma}{\sqrt{\widetilde{k}_n}}\left(r_n + \frac{3}{2}\sqrt{8\ln(2/\delta)} + \frac{3\ln(2/\delta)}{2\sqrt{\widetilde{k}_n}}\right)\,.
\end{aligned}$$

Therefore, plugging this bound into (4.6), with probability larger than $1-4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n)-\gamma\right| \leq \frac{\gamma}{\sqrt{\widetilde{k}_n}}\left(\left(r_n+\frac{3}{2}\sqrt{8\ln(2/\delta)}+\frac{3\ln(2/\delta)}{2\sqrt{\widetilde{k}_n}}\right)\left(1+\frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\right)+r_n(\delta)\right) .$$

4.8. *Proof of Corollary 3.18*

If, for some $C>0$ and $\rho<0$,

$$\left|\gamma-\mathbb{E}\widehat{\gamma}(k)\right| \leq C\left(\frac{n}{k}\right)^{\rho} ,$$

then

$$\frac{\gamma r_n}{\sqrt{\widetilde{k}_n+1}} \leq C\left(\frac{n}{\widetilde{k}_n+1}\right)^{\rho} ,$$

that is,

$$\sqrt{\widetilde{k}_n+1} \geq \left(\frac{\gamma r_n}{C}\right)^{1/(1+2|\rho|)} n^{|\rho|/(1+2|\rho|)} .$$

Thus, there exists a constant c such that

$$\sqrt{\widetilde{k}_n} \geq \left(\frac{\gamma r_n}{c}\right)^{1/(1+2|\rho|)} n^{|\rho|/(1+2|\rho|)} .$$

Starting from Equation (3.15) of Theorem 3.14, with probability $1-4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n)-\gamma\right| \leq 16\gamma\sqrt{\frac{2\ln(2/\delta\ln n)}{\widetilde{k}_n}}\left(1+\alpha(\delta,n)\right) ,$$

and, there exists a constant $\kappa_{\delta,\rho}$, depending on δ and ρ, such that

$$\sqrt{\frac{\ln(2/\delta\ln n)}{\widetilde{k}_n}} \leq \kappa_{\delta,\rho}\gamma^{-1/(1+2|\rho|)}\left(\frac{\ln(2/\delta\ln n)}{n}\right)^{|\rho|/(1+2|\rho|)} .$$

Hence, with probability larger than $1-4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n)-\gamma\right| \leq \kappa_{\delta,\rho}\left(\frac{\gamma^2\ln(2/\delta\ln n)}{n}\right)^{|\rho|/(1+2|\rho|)}\left(1+\alpha(\delta,n)\right) .$$

5. Simulations

Risk bounds like Theorem 3.14 and Corollary 3.18 are conservative. For all practical purposes, they are just meant to be reassuring guidelines. In this numerical section, we intend to shed some light on the following issues:

1. Is there a reasonable way to calibrate the threshold $r_n(\delta)$ used in the definition of $\widehat{k}_n$? How does the method perform if we choose $r_n(\delta)$ close to $\sqrt{2\ln\ln(n)}$?
2. How large is the ratio between the risk of $\widehat{\gamma}(\widehat{k}_n)$ and the risk of $\widehat{\gamma}(k_n^*)$ for moderate sample sizes?

The finite-sample performance of the data-driven index selection method described and analysed in Section 3.3 has been assessed by Monte-Carlo simulations. Computations have been carried out in `R` using packages `ggplot2`, `knitr`, `foreach`, `iterators`, `xtable` and `dplyr` [See Wickham, 2014, for a modern account of the R environment]. To get into the details, we investigated the performance of index selection methods on samples of sizes $10000, 20000$ and 100000 from the collection of distributions listed in Table 1. The list comprises the following distributions.

i) Fréchet distributions $F_\gamma(x) = \exp(x^{-1/\gamma})$ for $x > 0$ and $\gamma \in \{1, 0.5, 0.2\}$.
ii) Student distributions t_ν with $\nu \in \{1, 2, 4, 10\}$ degrees of freedom.
iii) log-gamma distribution with density proportional to $(\ln(x))^{2-1}x^{-3-1}$, which means $\gamma = 1/3$ and $\rho = 0$.
iv) The Lévy distribution with density $\sqrt{1/(2\pi)}\frac{e^{-\frac{1}{2x}}}{x^{3/2}}$, $\gamma = 2$ and $\rho = -1$ (this is the distribution of $1/X^2$ when $X \sim \mathcal{N}(0,1)$).
v) The H distribution is defined by $\gamma = 1/2$ and von Mises function equal to $\eta(s) = (2/s)\ln 1/s$. This distribution satisfies the second-order regular variation condition with $\rho = -1$ but does not fit into Model (2.7).
vi) Two Pareto change point distributions with distribution functions

$$\overline{F}(x) = x^{-1/\gamma'}\mathbb{1}_{\{1\le x\le\tau\}} + \tau^{-1/\gamma'}(x/\tau)^{-1/\gamma}\mathbb{1}_{\{x\ge\tau\}}$$

and $\gamma \in \{1.5, 1.25\}$, $\gamma' = 1$, and thresholds τ adjusted in such a way that they respectively correspond to quantiles of order $1 - 1/15$ and $1 - 1/25$.

Fréchet, Student, log-gamma distributions were used as benchmarks by [Drees and Kaufmann, 1998], [Danielsson et al., 2001] and [Carpentier and Kim, 2014a].

Table 1, which is complemented by Figure 3, describes the difficulty of tail index estimation from samples of the different distributions. Monte-Carlo estimates of the standardised root mean square error (RMSE) of Hill estimators

$$\mathbb{E}\left[\left(\widehat{\gamma}(k)/\gamma - 1\right)^2\right]^{1/2}$$

are represented as functions of the number of order statistics k for samples of size 10000 from the sampling distributions. All curves exhibit a common pattern: for small values of k, the RMSE is dominated by the variance term and scales like $1/\sqrt{k}$. Above a threshold that depends on the sampling distribution but that is not completely characterised by the second-order regular variation index, the RMSE grows at a rate that may reflect the second-order regular variation property of the distribution. Not too surprisingly, the three Fréchet distributions exhibit the same risk profile. The three curves are almost undistinguishable. The

Student distributions illustrate the impact of the second-order parameter on the difficulty of the index selection problem. For sample size $n = 10000$, the optimal index for t_{10} is smaller than 30, it is smaller than the usual recommendations and for such moderate sample sizes seems as hard to handle as the log-gamma distribution which usually fits in the Horror Hill Plot gallery. The 1/2-stable Lévy distribution and the H-distribution behave very differently. Even though they both have second-order parameter ρ equal to -1, the H distribution seems almost as challenging as the t_4 distribution while the Lévy distribution looks much easier than the Fréchet distributions. The Pareto change point distributions exhibit an abrupt transition.

TABLE 1
Estimated oracle index k_n^* and standardised RMSE $\mathbb{E}[(\gamma - \widehat{\gamma}(k_n^*))^2]^{1/2}/\gamma$ *for benchmark distributions. Estimates were computed from* 5000 *replicated experiments on samples of size* 10000.

d.f.	γ	ρ	k_n^*	RMSE
$F_{0.2}$	0.2	1.0	1132	3.7e-02
$F_{0.5}$	0.5	1.0	1145	3.6e-02
F_1	1.0	1.0	1155	3.6e-02
t_1	1.0	2.0	1161	3.3e-02
t_2	0.5	1.0	341	6.5e-02
t_4	0.2	0.5	77	1.6e-01
t_{10}	0.1	0.2	15	5.3e-01
H	0.5	1.0	130	1.1e-01
log-gamma	0.3	0.0	213	1.6e-01
Stable	2.0	1.0	3172	2.0e-02
Pcp	1.5	0.3	943	3.3e-02
Pcp (bis)	1.2	0.2	593	4.2e-02

Index $\widehat{k}_n(r_n)$ was computed according to the following rule

$$\widehat{k}_n(r_n) = \min\left\{ k\colon 30 \le k \le n \text{ and } \exists i \in \{30,\dots,k\}\,, \left|\widehat{\gamma}(i) - \widehat{\gamma}(k)\right| > \frac{r_n \widehat{\gamma}(i)}{\sqrt{i}} \right\} - 1 \qquad (5.1)$$

with $r_n = \sqrt{c \ln\ln n}$ where $c = 2.1$ unless otherwise specified.

The Fréchet, Student, H and stable distributions all fit into the framework considered by [Drees and Kaufmann, 1998]. They provide a favorable ground for comparing the performance of the optimal index selection method described by Drees and Kaufmann [1998] which attempts to take advantage of the second-order regular variation property and the performance of the simple selection rule described in this paper.

Index $\widehat{\gamma}(\widehat{k}_n^{\mathrm{DK}})$ was computed following the recommandations from Theorem 1 and discussion in [Drees and Kaufmann, 1998]:

$$\widehat{k}_n^{\mathrm{DK}} = (2|\widehat{\rho}| + 1)^{-1/|\widehat{\rho}|} \left(2\widehat{\rho}\widehat{\gamma}\right)^{1/(1+2|\hat{\rho}|)} \left(\frac{\widehat{k}_n(r_n^\zeta)}{(\widehat{k}_n(r_n))^\zeta} \right)^{1/(1-\zeta)} \qquad (5.2)$$

where $\widehat{\rho}$ should belong to a consistent family of estimators of ρ (under a second-order regular variation assumption), $\widehat{\gamma}$ should be a preliminary estimator of γ

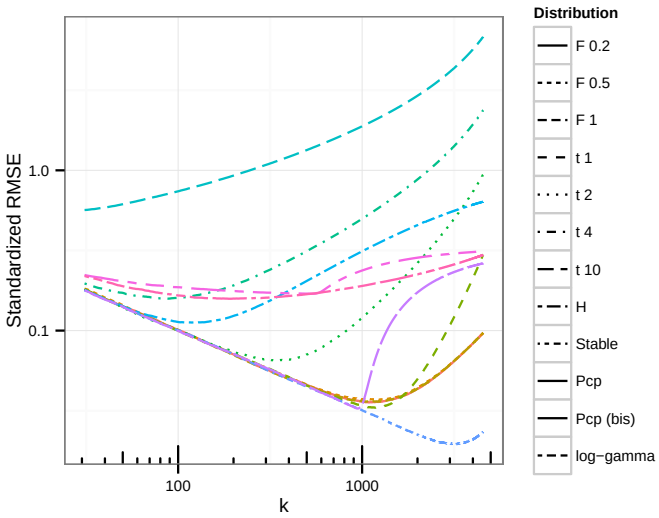

FIG 3. *Monte-Carlo estimates of the standardised root mean square error (*RMSE*) of Hill estimators as a function of the number of order statistics k for samples of size* 10000 *from the sampling distributions.*

such as $\widehat{\gamma}(\sqrt{n})$, $\zeta = .7$, and $r_n = 2n^{1/4}$. Following the advice from [Drees and Kaufmann, 1998], we replaced $|\widehat{\rho}|$ by 1. Note that the method for computing $\widehat{k}_n^{\mathrm{DK}}$ depends on a variety of tunable parameters.

Comparison between performances of $\widehat{\gamma}(\widehat{k}_n(r_n))$ and $\widehat{\gamma}(\widehat{k}_n^{\mathrm{DK}})$ are reported in Tables 2 and 3. For each distribution from Table 1, for sample sizes $n = 10000, 20000,$ and 1000000, 5000 experiments were replicated. As pointed out in [Drees and Kaufmann, 1998], on the sampling distributions that satisfy a second-order regular variation property, carefully tuned $\widehat{k}_n^{\mathrm{DK}}$ is able to take advantage of it. Despite its computational and conceptual simplicity, despite the fact that it is almost parameter free, the estimator $\widehat{\gamma}(\widehat{k}_n(r_n))$ only suffers a moderate loss with respect to the oracle. When $|\rho| = 1$, the observed ratios are of the same order as $(2 \ln \ln n)^{1/3} \approx 1.65$. Moreover, wheres $\widehat{\gamma}(\widehat{k}_n^{\mathrm{DK}})$ behaves erratically when facing Pareto change point distributions, $\widehat{\gamma}(\widehat{k}_n(r_n))$ behaves consistently.

Figure 4 concisely describes the behaviour of the two index selection methods on samples from the Pareto change point distribution with parameters $\gamma = 1.5, \gamma' = 1$ and threshold τ corresponding to the $1 - 1/15$ quantile. The plain line represents the standardised RMSE of Hill estimators as a function of selected index. This figure contains the superposition of two density plots corresponding to $\widehat{k}_n^{\mathrm{DK}}$ and $\widehat{k}(r_n)$. The density plots were generated from 5000 points with coordinates $(\widehat{k}(r_n), |\widehat{\gamma}(\widehat{k}(r_n))/\gamma - 1|)$ and 5000 points with coordi-

TABLE 2
Ratios between median selected indexes $\widehat{k}_n(r_n)$ (Lepski), $\widehat{k}_n^{\mathrm{DK}}$ (Drees-Kaufmann) and estimated oracle index k_n^*.

d.f.	γ	$\widehat{k}_n^{\mathrm{DK}}/k_n^*$			$\widehat{k}_n(r_n)/k_n^*$		
		$n = 1000$	20000	10000	10000	20000	100000
$F_{0.2}$	0.2	0.61	0.67	0.94	2.94	2.97	3.47
$F_{0.5}$	0.5	1.12	1.18	1.45	2.90	2.87	2.91
F_1	1	1.76	2.05	2.32	2.90	3.10	2.93
t_1	1	1.33	1.55	1.98	2.03	2.16	2.16
t_2	0.5	1.00	0.99	0.91	3.05	3.06	2.96
t_4	0.25	1.27	1.28	1.18	5.62	5.50	5.30
t_{10}	0.1	2.00	1.54	2.28	13.87	10.92	14.12
H	0.5	0.41	0.35	0.30	5.14	4.97	4.96
Stable	2	0.97	0.95	1.04	1.43	1.41	1.55
Pcp	1.5	1.85	0.45	0.15	1.32	1.21	1.10
Pcp (bis)	1.25	3.29	3.03	2.45	1.83	1.50	1.22
log-gamma	0.33	5.13	7.71	12.41	10.50	12.99	12.40

TABLE 3
Ratios between median RMSE *of and median optimal* RMSE.

d.f.	γ	$\mathrm{RMSE}(\widehat{\gamma}(\widehat{k}_n^{\mathrm{DK}}))/\mathrm{RMSE}(\widehat{\gamma}(k_n^*))$			$\mathrm{RMSE}(\widehat{\gamma}(\widehat{k}_n(r_n)))/\mathrm{RMSE}(\widehat{\gamma}(k_n^*))$		
		$n = 1000$	20000	10000	10000	20000	100000
$F_{0.2}$	0.2	1.12	1.12	1.02	2.06	2.26	2.69
$F_{0.5}$	0.5	1.03	1.03	1.14	2.12	2.23	2.70
F_1	1	1.22	1.31	1.59	2.07	2.23	2.64
t_1	1	1.26	1.34	1.74	2.31	2.39	3.11
t_2	0.5	1.11	1.08	1.05	2.06	2.09	2.20
t_4	0.25	1.10	1.07	1.04	1.85	1.81	1.84
t_{10}	0.1	1.10	1.09	1.08	1.76	1.72	1.64
H	0.5	1.28	1.37	1.48	2.15	2.18	2.12
Stable	2	1.01	0.99	0.98	1.99	2.52	3.60
Pcp	1.5	4.25	1.66	2.52	2.50	2.68	3.63
Pcp (bis)	1.25	3.38	4.47	7.45	2.43	2.56	3.10
log-gamma	0.33	1.23	1.28	1.39	1.45	1.43	1.37

nates $(\widehat{k}_n^{\mathrm{DK}}, |\widehat{\gamma}(\widehat{k}_n^{\mathrm{DK}})/\gamma - 1|)$. The contoured and well-concentrated density plot corresponds to the performance of $\widehat{\gamma}(\widehat{k}_n)$. The diffuse tiled density plot corresponds to the performance of $\widehat{k}_n^{\mathrm{DK}}$. Facing Pareto change point samples, the two selection methods behave differently. Lepski's rule detects correctly an abrupt change at some point and selects an index slightly above that point. As the conditional bias varies sharply around the change point, this slight over estimation of the correct index still results in a significant loss as far as RMSE is concerned. The Drees-Kaufmann rule, fed with an a priori estimate of the second-order parameter, picks out a much smaller index, and suffers a larger excess risk.

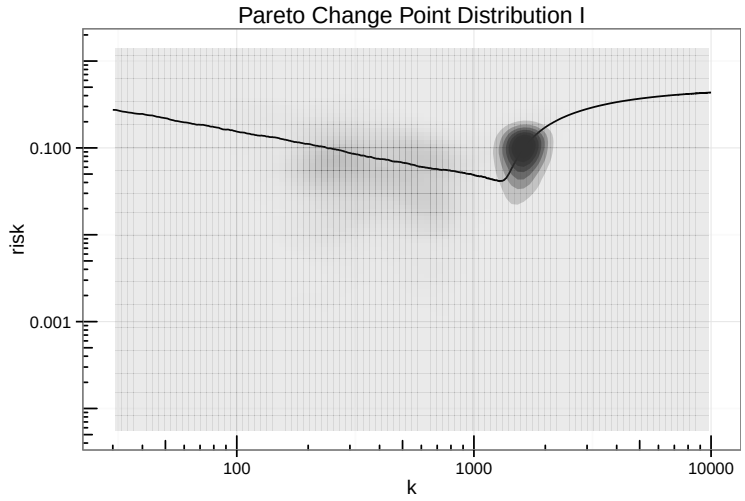

FIG 4. *Risk plot for samples of size* 10000 *from the Pareto change point distribution with parameters* $\gamma = 1.5, \gamma' = 1$ *and threshold* τ *corresponding to the* $1 - 1/15$ *quantile. The concentrated density plot corresponds to points* $(\widehat{k}(r_n), |\widehat{\gamma}(\widehat{k}(r_n))/\gamma - 1|)$.

References

J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2004.

J. Beirlant, C. Bouquiaux, and B. Werker. Semiparametric lower bounds for tail index estimation. *Journal of Statistical Planning and Inference*, 136(3): 705–729, 2006.

N. Bingham, C. Goldie, and J. Teugels. *Regular variation*. Cambridge University Press, 1987.

L. Birgé. An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist., Beachwood, OH, 2001.

L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51:1611–1615, 2005.

S. Bobkov and M. Ledoux. Poincaré's inequalities and Talagrands's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107:383–400, 1997.

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electron. Commun. Probab.*, 17:1–12, 2012. ISSN 1083-589X. . URL http://dx.doi.org/10.1214/ECP.v17-2210.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, 2013.

A. Carpentier and A. Kim. Adaptive and minimax optimal estimation of the tail coefficient. *Preprint available at http://arxiv.org/pdf/1309.2585.pdf*, 2014a.

A. Carpentier and A. Kim. Adaptive confidence intervals for the tail coefficient in a wide second order class of pareto models. *Preprint available at http://arxiv.org/pdf/1312.2968v2.pdf*, 2014b.

S. Chatterjee. *Superconcentration and related topics*. Springer, Cham, 2014.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

S. Csörgő, P. Deheuvels, and D. Mason. Kernel estimates of the tail index of a distribution. *Ann. Statist.*, 13(3):1050–1077, 1985.

J. Danielsson, L. de Haan, L. Peng, and C. G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.*, 76(2):226–248, 2001. ISSN 0047-259X. . URL http://dx.doi.org/10.1006/jmva.2000.1903.

D. Darling and P. Erdös. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J*, 23:143–155, 1956.

L. de Haan and A. Ferreira. *Extreme value theory*. Springer-Verlag, 2006.

G. Draisma, L. de Haan, L. Peng, and T. Pereira. A bootstrap-based method to achieve optimally in estimating the extreme value index. *Extremes*, 2:367–404, 1999.

H. Drees. Optimal rates of convergence for estimates of the extreme value index. *Annals of Statistics*, 26(1):434–448, 1998a.

H. Drees. On smooth statistical tail functionals. *Scandinavian Journal of Statistics*, 25(1):187–210, 1998b. ISSN 1467-9469.

H. Drees. Minimax risk bounds in extreme value theory. *Ann. Statist.*, 29(1): 266–294, 2001. ISSN 0090-5364. . URL http://dx.doi.org/10.1214/aos/996986509.

H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.*, 75(2):149–172, 1998. ISSN 0304-4149. . URL http://dx.doi.org/10.1016/S0304-4149(98)00017-9.

H. Drees, L. De Haan, and S. Resnick. How to make a Hill plot. *Annals of Statistics*, 28(1):254–274, 2000.

J. Geluk, L. de Haan, S. Resnick, and C. Stărică. Second-order regular variation, convolution and the central limit theorem. *Stochastic Processes and Applications Appl.*, 69(2):139–159, 1997.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216, 2006.

I. Grama and V. Spokoiny. Statistics of extremes by oracle estimation. *Ann. Statist.*, 36(4):1619–1648, 2008. ISSN 0090-5364. . URL http://dx.doi.org/10.1214/07-AOS535.

P. Hall and I. Weissman. On the estimation of extreme tail probabilities. *Ann. Statist.*, 25(3):1311–1326, 1997.

P. Hall and A. Welsh. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13(1):331–341, 1985.

B. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3:1163–1174, 1975.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems. Lectures from the 38th Probability Summer School, Saint-Flour*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2008.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

O. Lepski. A problem of adaptive estimation in Gaussian white noise. *Teoriya Veroyatnosteui i ee Primeneniya*, 35(3):459–470, 1990.

O. Lepski. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teoriya Veroyatnosteui i ee Primeneniya*, 36(4): 645–659, 1991.

O. Lepski. Asymptotically minimax adaptive estimation. II. schemes without optimal adaptation. adaptive estimates. *Teoriya Veroyatnosteui i ee Primeneniya*, 37(3):468–481, 1992.

O. Lepski and A. Tsybakov. Asymptotic exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*, 117 (1):17–48, 2000.

D. Mason. Laws of large numbers for sums of extreme values. *Annals of Probability*, 10:754–764, 1982.

P. Massart. *Concentration inequalities and model selection. Ecole d'Eté de Probabilité de Saint-Flour* XXXIV, volume 1896 of *Lecture Notes in Mathematics*. Springer-Verlag, 2007.

P. Mathé. The Lepski principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.

B. Maurey. Some deviation inequalities. *Geometric and Functional Analysis*, 1 (2):188–197, 1991.

S. Novak. Lower bounds to the accuracy of inference on heavy tails. *Bernoulli*, 20(2):979–989, 2014.

S. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*, volume 10. Springer Verlag, 2007.

J. Segers. Abelian and Tauberian theorems on the bias of the Hill estimator. *Scand. J. Statist.*, 29(3):461–483, 2002. ISSN 0303-6898. . URL http://dx.doi.org/10.1111/1467-9469.00301.

M. Talagrand. A new isoperimetric inequality and the concentration of measure phenomenon. In *Geometric aspects of functional analysis (1989–90)*, volume 1469 of *Lecture Notes in Math.*, pages 94–124. Springer, Berlin, 1991.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24:1–34, 1996a. (Special Invited Paper).

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996b.

M. Talagrand. *The generic chaining*. Springer, New York, 2005.

S. van de Geer. *Applications of empirical process theory*. Cambridge University Press, 2000.

H. Wickham. *Advanced R*. Chapman & Hall/CRC, 2014.

Appendix A: Calibration of the preliminary selection rule

Darling and Erdös [1956] establish (among other things) that letting Z_n denote $\sup_{k\leq n} \sum_{i=1}^{k}(E_i-1)/\sqrt{k}$, where E_i's are independent exponentially distributed random variables, the sequence $\sqrt{2\ln\ln n}\left(Z_n - \sqrt{2\ln\ln n - \ln\ln\ln(n)}\right)$ converges in distribution towards a non degenerate distribution which is closely related to the Gumbel distribution. In other words, asymptotically, Z_n behaves almost like the maximum of $\ln n$ independent standard Gaussian random variables.

Appendix B: Proof of Corollary 2.16

Let $Z = f(E_1,\ldots,E_n) = (U\circ\exp)\left(\sum_{i=1}^{k} E_i/i\right)$. Then,

$$|\partial_i f| \leq \frac{1}{i}\sup_x \frac{1}{h(x)}, \quad \text{for } i \geq k$$

and

$$\|\nabla f\|^2 = \sum_{i=k}^{n} \frac{1}{i^2}\frac{1}{(h\circ f)^2} \ .$$

Let $c<1$, then for all $\lambda, 0\leq\lambda\leq c\,(k\inf_x h(x))$,

$$\log \mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)} \leq \frac{4/k\,(1+1/k)\,\mathbb{E}[1/h(Z)^2]\lambda^2}{2(1-c)} \ .$$

Now, start from the first statement in Theorem 2.15,

$$\begin{aligned}
\mathrm{Ent}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right] &\leq \frac{2\lambda^2}{1-c}\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\|\nabla f\|^2\right] \\
&= \frac{4\lambda^2}{2(1-c)}\frac{1}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\frac{\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}}{h(Z)^2}\right] \\
&\leq \frac{4\lambda^2}{2(1-c)}\frac{1}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right]\mathbb{E}\left[\frac{1}{h(Z)^2}\right]
\end{aligned}$$

where the last inequality follows from Chebychev negative association inequality. Hence,

$$\frac{\mathrm{d}\log\mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}}{\mathrm{d}\lambda} = \frac{\mathrm{Ent}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right]}{\lambda^2\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right]} \leq \frac{1}{2(1-c)}\frac{4}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\frac{1}{h(Z)^2}\right] \ .$$

This differential inequality is readily solved and leads to the corollary.

Appendix C: Proof of Abelian Proposition 3.2

The proof proceeds by classical arguments. In the sequel, we use the almost sure representation argument. Without loss of generality, we assume that all the random variables live on the same probability space, and that for any intermediate sequence $(k_n)_n$, $\sqrt{k_n}(Y_{(k_n+1)} - \ln(n/k_n))$ converges almost surely towards a standard Gaussian random variable. Complemented with dominated convergence arguments, the next lemma will be the key element of the proof.

Lemma C.1. *Let $\eta \in \mathsf{RV}_\rho, \rho \leq 0$ and $Y_{(k_n+1)}$ be the (k_n+1)th largest order statistic of a standard exponential sample, then for any intermediate sequence $(k_n)_n$ and for all $u > 0$,*

$$\lim_{n\to\infty} \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)} = \mathrm{e}^{\rho u} \ p.s \ .$$

Proof. Note that

$$\frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)} \quad = \quad \frac{\eta((n/k_n)\mathrm{e}^{u+Y_{(k_n+1)}-\log(n/k_n)})}{\eta(n/k_n)} \, .$$

Then, the result follows since $Y_{(k_n+1)} - \log(n/k_n) \xrightarrow{\text{p.s}} 0$ and the convergence $\eta(tx)/\eta(t) \to x^\rho$ is locally uniform on $(0,\infty)$. □

In order to secure dominated convergence arguments, we will use Drees's improvement of Potter's inequality [See de Haan and Ferreira, 2006, page 369]. For every $\epsilon, \delta > 0$, there exists $t_0 = t_0(\epsilon,\delta)$ such that for $t, tx \geq t_0$,

$$|\eta(tx)/\eta(t) - x^\rho| \leq x^\rho \epsilon \max(x^\delta, x^{-\delta}) \, . \tag{C.2}$$

To prove Proposition 3.2, we start from Representation (2.4):

$$\widehat{\gamma}(k_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} \int_0^{E_i} \left(\gamma + \eta\left(\mathrm{e}^{u+Y_{(k_n+1)}}\right)\right) \mathrm{d}u \ .$$

By the Pythagorean relation,

$$\mathrm{Var}(\widehat{\gamma}(k_n)) = \mathrm{Var}\left(\mathbb{E}[\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}]\right) + \mathbb{E}\left[\mathrm{Var}\left(\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}\right)\right] \, ,$$

so that

$$\begin{aligned} &\frac{k_n \,\mathrm{Var}(\widehat{\gamma}(k_n)) - \gamma^2}{\eta\left(n/k_n\right)} \\ &= \quad \frac{k_n \,\mathrm{Var}\left(\mathbb{E}[\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}]\right)}{\eta(n/k_n)} + k_n \mathbb{E}\left[\frac{\mathrm{Var}\left(\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}\right) - \frac{\gamma^2}{k_n}}{\eta(n/k_n)}\right] . \end{aligned}$$

The second summand can be further decomposed using (2.4),

$$\begin{aligned}
&\frac{k_n \operatorname{Var}(\widehat{\gamma}(k_n)) - \gamma^2}{\eta\left(n/k_n\right)} \\
&= \underbrace{\frac{k_n \operatorname{Var}\left(\mathbb{E}[\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}]\right)}{\eta(n/k_n)}}_{\text{(I)}} + \underbrace{\eta(\tfrac{n}{k_n})\mathbb{E}\left[\operatorname{Var}\left[\int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u \mid Y_{(k_n+1)}\right]\right]}_{\text{(II)}} \\
&\quad + \underbrace{2\gamma\mathbb{E}\left[\operatorname{Cov}\left[E, \int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u \mid Y_{(k_n+1)}\right]\right]}_{\text{(III)}}.
\end{aligned}$$

We check that (I) and (II) tend to 0 and then that (III) converges towards a finite limit.

Fix $\epsilon, \delta > 0$ and define $M = \sup\{\eta(t), t \le t_0\}$.
Let A_n denote the event $\{Y_{(k_n+1)} > \ln t_0(\epsilon, \delta)\}$. For n such that $\ln(n/k_n) \le 2\ln t_0$, as $Y_{(k_n+1)}$ sub-gamma with variance factor $1/k_n$,

$$\mathbb{P}\{A_n^c\} \le \exp\left(-k_n(\ln(n/k_n))^2/8\right).$$

We first check that (II) tends to 0. Let n be such that $n/k_n \ge t_0$ and W_n denote the random variable $Y_{(k_n+1)} - \ln(n/k_n)$. Note that for $0 \le \lambda \le k_n/2$,

$$\mathbb{E}\mathrm{e}^{\lambda|W_n|} \le 2\mathrm{e}^{\frac{\lambda^2}{k_n}}.$$

Using Jensen's inequality and Fubini's Theorem,

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Var}\left[\int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u \mid Y_{(k_n+1)}\right]\right] &\le \mathbb{E}\left[\mathbb{E}\left[E\int_0^E \left(\frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\right)^2 \mathrm{d}u \mid Y_{(k_n+1)}\right]\right] \\
&= \int_0^\infty \mathrm{e}^{-v} v \int_0^v \mathbb{E}\left[\left(\frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\right)^2\right]\mathrm{d}u\mathrm{d}v \\
&= \int_0^\infty \mathrm{e}^{-v} v \int_0^v \mathbb{E}\left[\left(\frac{\eta(\mathrm{e}^{u+W_n} n/k_n)}{\eta(n/k_n)}\right)^2\right]\mathrm{d}u\mathrm{d}v
\end{aligned}$$

We now apply Potter's inequality (C.2) on the event A_n with $t = n/k_n > t_0$ and $tx = \mathrm{e}^{u+Y_{(k_n+1)}} > t_0, u > 0$:

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Var}\left[\int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u \mid Y_{(k_n+1)}\right]\right] \\
&\le \int_0^\infty \mathrm{e}^{-v} v \int_0^v \mathbb{E}\left[\mathbb{1}_{A_n}\mathrm{e}^{2\rho(u+W_n)}\left(1+\epsilon\mathrm{e}^{\delta(u+|W_n|)}\right)^2 + \mathbb{1}_{A_n^c}\frac{M^2}{\eta(n/k_n)^2}\right]\mathrm{d}u\mathrm{d}v \\
&\le \int_0^\infty \mathrm{e}^{-v} v \int_0^v \mathbb{E}\left[\mathrm{e}^{2\rho W_n} 2\left(1+\epsilon^2\mathrm{e}^{2\delta(u+|W_n|)}\right)\right]\mathrm{d}u\mathrm{d}v + \frac{2M^2}{\eta(n/k_n)^2}\mathbb{E}\mathbb{1}_{A_n^c}.
\end{aligned}$$

The first summand has a finite limit thanks to Lemma C.1. The second summand converges to 0 as $\mathbb{E}\mathbb{1}_{A_n^c}$ tends to 0 exponentially fast while $1/\eta(n/k_n)^2$ tends to infinity algebraically fast.

Bounds on (I) are easily obtained, using Jensen's Inequality and Poincaré Inequality.

$$\begin{aligned}\frac{k_n \operatorname{Var}\left(\mathbb{E}[\widehat{\gamma}(k_n) \mid Y_{(k_n+1)}]\right)}{\eta(n/k_n)} &= \frac{k_n \operatorname{Var}\left(\int_0^\infty \eta\left(\mathrm{e}^{u+Y_{(k_n+1)}}\right)\mathrm{e}^{-u}\mathrm{d}u\right)}{\eta(n/k_n)} \\ &\leq 4\eta(n/k_n)\mathbb{E}\left[\left(\int_0^\infty \frac{\eta\left(\mathrm{e}^{u+Y_{(k_n+1)}}\right)}{\eta(n/k_n)}\mathrm{e}^{-u}\mathrm{d}u\right)^2\right] \\ &\leq 4\eta(n/k_n)\mathbb{E}\left[\int_0^\infty \left(\frac{\eta\left(\mathrm{e}^{u+Y_{(k_n+1)}}\right)}{\eta(n/k_n)}\right)^2 \mathrm{e}^{-u}\mathrm{d}u\right].\end{aligned}$$

Using the line of arguments as for handling the limit of (II), we establish that (I) converges to 0.

We now check that (III) converges towards a finite limit. Note that

$$\mathbb{E}\left[\operatorname{Cov}\left[E, \int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u \mid Y_{(k_n+1)}\right]\right] = \mathbb{E}\left[(E-1)\int_0^E \frac{\eta(\mathrm{e}^{u+Y_{(k_n+1)}})}{\eta(n/k_n)}\mathrm{d}u\right].$$

By Lemma C.1, for almost every $u > 0$,

$$(E-1)\frac{\eta(\mathrm{e}^{u+W_n}n/k_n)}{\eta(n/k_n)} \underset{n\to\infty}{\longrightarrow} (E-1)\mathrm{e}^{\rho u},$$

and

$$|E-1|\int_0^E \left|\frac{\eta(\mathrm{e}^{u+W_n}n/k_n)}{\eta(n/k_n)}\right|\mathrm{d}u \leq |E-1|\int_0^E \mathrm{e}^{\rho(u+W_n)}\left(1+\epsilon\mathrm{e}^{\delta(u+|W_n|)}\right)\mathrm{d}u + \mathbb{1}_{A_n^c}E|E-1|\frac{M}{|\eta(n/k_n)|}.$$

The first term is finite as the integral of a continuous function on a compact. Thus,

$$(E-1)\int_0^E \frac{\eta(\mathrm{e}^{u+W_n}n/k_n)}{\eta(n/k_n)}\mathrm{d}u \to_n (E-1)\int_0^E \mathrm{e}^{\rho u}\mathrm{d}u = (E-1)\frac{\mathrm{e}^{\rho E}-1}{\rho}.$$

The expected value of the last random variable is $1/(1-\rho)^2$.

We check that for sufficiently large n,

$$\begin{aligned}&\mathbb{E}\left[|E-1|\int_0^E \frac{|\eta(\mathrm{e}^{u+W_n}n/k_n)|}{|\eta(n/k_n)|}\mathrm{d}u\right] \\ &\leq \mathbb{E}\left[|E-1|\int_0^E \mathrm{e}^{\rho(u+W_n)}\left(1+\epsilon\mathrm{e}^{\delta(u+|W_n|)}\right) + \mathbb{1}_{A_n^c}|E-1|\frac{M}{|\eta(n/k_n)|}\mathrm{d}u\right] \\ &\leq \mathbb{E}\left[\mathrm{e}^{\rho W_n}\left(2+\frac{\epsilon}{\delta(1-\delta)^2}\mathrm{e}^{\delta|W_n|}\right)\right] + \frac{M}{|\eta(n/k_n)|}\mathbb{E}\mathbb{1}_{A_n^c} \\ &\leq 4\mathrm{e}^{\frac{\rho^2}{k_n}} + \frac{2\epsilon}{\delta(1-\delta)^2}\mathrm{e}^{\frac{(\delta-\rho)^2}{k_n}} + \frac{M}{|\eta(n/k_n)|}\mathbb{E}\mathbb{1}_{A_n^c}.\end{aligned}$$

We now way conclude by dominated convergence that

$$(\text{III}) \underset{n\to\infty}{\longrightarrow} \frac{2\gamma}{(1-\rho)^2} \ .$$

Appendix D: Revisiting the lower bound on adaptive estimation error

Lower bounds on tail index estimation error [Carpentier and Kim, 2014a, Drees, 1998a, 2001, Novak, 2014] are usually constructed by defining sequences of local models around a pure Pareto distribution with shape parameter γ_0. When deriving lower bounds for the estimation error under constraints like $\overline{\eta}$ is regularly varying, the elements of the local model for sample size n may be defined by

$$U_{n,h}(t) = t^{\gamma_0 + d_n h(0)} \exp \int_1^t d_n \frac{h(c_n/s) - h(0)}{s} \mathrm{d}s$$

where h is square integrable over $[0,1]$, $d_n \to 0$, $nd_n^2/c_n \to 1$ [Drees, 2001]. The sequences d_n and c_n are chosen in such a way that $d_n \left|h(c_n/s) - h(0)\right| = \left|\overline{\eta}(s)\right|$ satisfies the required constraint. If the local alternatives are Pareto change point distributions as in [Novak, 2014] and [Carpentier and Kim, 2014a], $h(x) = \mathbb{1}_{\{x \le 1\}}$, $c_n = \tau_n^{1/\gamma_0}$. Drees [2001] explores a richer collection of local alternatives in order to fit into the theory of weak convergence of local experiments.

In order to explore adaptivity as in [Carpentier and Kim, 2014a], it is necessary to handle simultaneously a collection of sequences $(d_n, c_n)_n$ corresponding to different rates of decay of the von Mises function. The difficulty of estimation is connected with the difficulty of distinguishing alternatives with different tail indexes that is, with the hardness of a multiple hypotheses testing problem. In order to lower bound the testing error, Carpentier and Kim chose to use Fano's Lemma [Cover and Thomas, 1991, See]. Using Fano's Lemma requires bounding the Kullback-Leibler divergence between the different local alternatives which is not as easy as bounding the divergence between a Pareto change point distribution and a pure Pareto distribution.

The next lemma is from [Birgé, 2005]. It can be used in the derivation of risk lower bounds instead of the classical Fano Lemma. Just as Fano's Lemma, it states a lower bound on the error in multiple hypothesis testing. But as it only requires computing the Kullback-Leibler divergence to the localisation center, in the present setting, it significantly alleviates computations and makes the proof more concise and more transparent.

Lemma D.1. *(Birgé-Fano) Let $P_0, \ldots, P_N$ be a collection of probability distributions on some space, and let $A_0, \ldots, A_M$ be a collection of pairwise disjoint events, then the following holds*

$$\min_i P_i\{A_i\} \le \frac{2\mathrm{e}}{1+2\mathrm{e}} \vee \frac{\frac{1}{M}\sum_{i=1}^M \mathcal{K}(P_i, P_0)}{\ln(M+1)} \ .$$

In order to take advantage of Lemma D.1, we use the Bayesian game designed in [Carpentier and Kim, 2014a].

Theorem D.2. *Let $\gamma > 0$, $\rho < -1$, and $0 \leq v \leq \mathrm{e}/(1+2\mathrm{e})$. Then, for any tail index estimator $\widehat{\gamma}$ and any sample size n such that $M = \lfloor \ln n \rfloor > \mathrm{e}/v$, there exists a collection $(P_i)_{i \leq M}$ of probability distributions such that*

i) $P_i \in \mathit{MDA}(\gamma_i)$ with $\gamma_i > \gamma$,

ii) P_i meets the von Mises condition with von Mises function η_i satisfying

$$\overline{\eta}_i(t) \leq \gamma t^{\rho_i}$$

where $\rho_i = \rho + i/M < 0$,

iii)

$$\max_{i \leq M} P_i^{\otimes n} \left\{ |\widehat{\gamma} - \gamma_i| \geq \frac{C_\rho}{4} \gamma_i \left(\frac{v \ln \ln n}{n} \right)^{|\rho_i|/(1+2|\rho_i|)} \right\} \geq \frac{1}{1+2\mathrm{e}}$$

and

$$\max_{i \leq M} \mathbb{E}_{P_i^{\otimes n}} \left[\frac{|\widehat{\gamma} - \gamma_i|}{\gamma_i} \right] \geq \frac{C_\rho}{4(1+2\mathrm{e})} \left(\frac{v \ln \ln n}{n} \right)^{|\rho|/(1+2|\rho|)},$$

with $C_\rho = 1 - \exp\left(-\frac{1}{2(1+2|\rho|)^2}\right)$.

Proof of Theorem D.2. Choose v so that $0 \leq v \leq 2\mathrm{e}/(1+2\mathrm{e})$. The number of alternative hypotheses M is chosen in such a way that $\ln\left(n/(v \ln M)\right) \leq M$. If $\lfloor \ln n \rfloor \geq \mathrm{e}/v$, $M = \lfloor \ln n \rfloor$ will do.

The center of localisation P_0 is the pure Pareto distribution with shape parameter $\gamma > 0$ ($P_0\{(\tau, \infty)\} = \tau^{-1/\gamma}$). The local alternatives $P_1, \ldots, P_M$ are Pareto change point distributions. Each P_i is defined by a breakpoint $\tau_i > 1$ and an ultimate Pareto index γ_i. If F_i denotes the distribution function of P_i,

$$\overline{F}_i(x) = x^{-1/\gamma} \mathbb{1}_{\{1 \leq x \leq \tau_i\}} + \tau_i^{-1/\gamma} (x/\tau_i)^{-1/\gamma_i} \mathbb{1}_{\{x \geq \tau_i\}}.$$

Karamata's representation of $(1/\overline{F}_i)^{\leftarrow}$ is

$$U_i(t) = t^{\gamma_i} \exp\left(\int_1^t \frac{\eta_i(s)}{s} \right) \mathrm{d}s$$

with $\eta_i(s) = (\gamma - \gamma_i) \mathbb{1}_{\{s \leq \tau_i^{1/\gamma}\}}$.

The Kullback-Leibler divergence between P_i and P_0 is readily calculated,

$$\mathcal{K}(P_i, P_0) = \overline{F}_i(\tau_i) \left(\frac{\gamma_i}{\gamma} - 1 - \ln \frac{\gamma_i}{\gamma} \right) = \tau_i^{-1/\gamma} \left(\frac{\gamma_i}{\gamma} - 1 - \ln \frac{\gamma_i}{\gamma} \right).$$

If $\gamma_i > \gamma$, the next upper bound holds,

$$\mathcal{K}(P_i, P_0) \leq \frac{\tau_i^{-1/\gamma}}{2} \left(\frac{\gamma_i}{\gamma} - 1 \right)^2.$$

The breakpoints and tail indices are chosen in such a way that all upper bounds are equal (namely $n\tau_i^{-1/\gamma}(\gamma_i/\gamma - 1)^2$ does not depend on i),

$$\begin{aligned} \tau_i &= (n/(v \ln M))^{\gamma/(1+2|\rho_i|)} \\ \gamma_i &= \gamma + \gamma\,(n/(v \ln M))^{\rho_i/(1+2|\rho_i|)}\,, \end{aligned}$$

so that $\mathcal{K}(P_i^{\otimes n}, P_0^{\otimes n}) = n\mathcal{K}(P_i, P_0) \le v \ln M$ for all $1 \le i \le M$.

Note that for all $t > 1$,

$$|\eta_i(t)| = |\gamma - \gamma_i| \mathbb{1}_{\{t \le \tau_i^{1/\gamma}\}} \le \gamma \tau_i^{\rho_i/\gamma} \mathbb{1}_{\{t \le \tau_i^{1/\gamma}\}} \le \gamma t^{\rho_i}$$

the upper bound being achieved at $t = \tau_i$.

Now, let $\widehat{\gamma}$ be any tail index estimator. Define region A_i, as the set of samples such that γ_i minimises $|\widehat{\gamma} - \gamma_j|$ for $1 \le j \le M$. Then, if the event A_i is not realised,

$$|\widehat{\gamma} - \gamma_i| \ge \frac{1}{2} \min_{1 \le j \le M, j \ne i} |\gamma_j - \gamma_i|\,.$$

By Birgé's Lemma,

$$\max_{i \le M} \mathbb{P}_i^{\otimes n} \left\{ |\widehat{\gamma} - \gamma_i| \ge \frac{1}{2} \min_{1 \le j \le M, j \ne i} |\gamma_j - \gamma_i| \right\} \ge \frac{1}{1 + 2\mathrm{e}}\,.$$

In order to make the whole construction useful, it remains to choose the "second-order parameters" ρ_i's (the true second-order parameter of each P_i is infinite!). We will need an upper bound on γ_i/γ (but we already have $\gamma_i/\gamma \le 2$), as well as a lower bound on $|\gamma_j - \gamma_i|/\gamma$ for $j \ne i$ that scales like $(n/\ln\ln n)^{\rho_i/(1+2|\rho_i|)}$.

Following Carpentier and Kim [2014a], we finally choose ρ_i as $\rho_i = \rho + i/M$ for $1 \le i \le M$. Then, for $j < i$, using that $\ln(n/(v \ln M)) \le M$ and $\rho_i - \rho_j = (i-j)/M$,

$$\begin{aligned} \frac{|\gamma_j - \gamma_i|}{\gamma_i} &\ge \frac{|\gamma_j - \gamma_i|}{2\gamma} \\ &\ge \frac{1}{2} \Big(\frac{n}{v \ln M}\Big)^{\rho_i/(1+2|\rho_i|)} \left| 1 - \Big(\frac{n}{v \ln M}\Big)^{\rho_j/(1+2|\rho_j|) - \rho_i/(1+2|\rho_i|)} \right| \\ &\ge \frac{1}{2} \Big(\frac{n}{v \ln M}\Big)^{\rho_i/(1+2|\rho_i|)} \left[1 - \exp\left(\frac{i-j}{M(1+2|\rho_i|)(1+2|\rho_j|)} \ln\Big(\frac{n}{v \ln M}\Big) \right) \right] \\ &\ge \frac{1}{2} \Big(\frac{n}{v \ln M}\Big)^{\rho_i/(1+2|\rho_i|)} \left[1 - \exp\left(\frac{i-j}{M(1+2|\rho_i|)(1+2|\rho_j|)} \right) \right] \\ &\ge \frac{C_\rho}{2} \Big(\frac{n}{v \ln M}\Big)^{\rho_i/(1+2|\rho_i|)} \end{aligned}$$

where C_ρ may be chosen as $1 - \exp\left(-\frac{1}{2(1+2|\rho|)^2}\right)$. □